Philosophy of Science Assoc. 24th Biennial Mtg

Chicago, IL

Version: 7 November 2014

PhilSci ARCHIVE

Table of Contents

Coherence, Muller's Ratchet, and the Maintenance of Culture

(Accepted for publication: A shortened version will appear in *Philosophy of Science* 82(5), 2015.)

Marshall Abrams (mabrams@uab.edu)

Department of Philosophy

University of Alabama at Birmingham

July 23, 2014

Abstract

I investigate the structure of an argument that culture cannot be maintained in a population if each individual acquires a given cultural variant from a single person. I note two puzzling consequences of the argument: It appears to conflict with (a) many models of cultural transmission and (b) real-world cases of cultural transmission. I resolve the first puzzle by showing that one of the models central to the argument is conceptually analogous and mathematically equivalent to one used to investigate the evolution of sexual reproduction. This analogy clarifies what assumptions are crucial to the argument concerning cultural transmission. I resolve the second puzzle by arguing that probabilistic models of epistemological coherence can be reinterpreted as models of mutual support between cultural variants. I develop a model of cultural transmission illustrating this proposal. I suggest that real-world cases that seem to conflict with the original argument may in fact be instances in which mutually supporting cultural variants are learned from different individuals.

1 Introduction

I investigate the structure of an argument that culture cannot be maintained in a population if each individual acquires a given cultural variant from a single person. I note two puzzling consequences of the argument: It appears to conflict with (a) many models of cultural transmission and (b) real-world cases of cultural transmission. I resolve the first puzzle by showing that one of the models central to the argument is conceptually analogous and mathematically equivalent to one used to investigate the evolution of sexual reproduction. This analogy clarifies what assumptions are crucial to the argument concerning cultural transmission. I resolve the second puzzle by arguing that probabilistic models of epistemological coherence can be reinterpreted as models of mutual support between cultural variants. I develop a model of cultural transmission illustrating this proposal. I suggest that real-world cases that seem to conflict with the original argument may in fact be instances in which mutually supporting cultural variants are learned from different individuals.

Section 2 describes arguments based on models of cultural transmission in (Enquist et al., 2010), and section 3 describes two implications of Enquist et al.'s results that may seem puzzling. Section 4 explains why the evolution of sexual reproduction is an interesting problem in evolutionary biology, and describes Muller's ratchet, a kind of model that motivates one of the proposed evolutionary advantages of sexual reproduction. This is the model that bears mathematical and conceptual parallels to one of Enquist et al.'s models. Section 5 then reinterprets ideas from probabilistic models of coherence in epistemology as ideas about cognitive transitions from one cultural variant to another.

I'll assume that there are at least minor differences in beliefs, knowledge, behavior, attitudes, inclinations, etc. between people within a society, and that such char-

acteristics—"cultural variants"—in one person sometimes promote similar variants in another.[1] Such processes are called "cultural transmission" or "social learning". "Individual learning", by contrast, occurs when a person learns something on their own, whether through observation of the environment, trial and error, reasoning, or some combination.

2 One Cultural Parent Makes No Culture

In this section I summarize arguments made by Enquist, Strimling, Eriksson, Laland, Sjostrand (2010) (henceforth "ESELS"). These authors argue that if a particular cultural variant is always acquired from a single individual—a single "cultural parent"—then it's very difficult for the cultural variant to be maintained in a population. Culture would usually disappear. Thus the maintenance of culture in human populations seems to depend on the fact that we routinely learn from multiple individuals.

Suppose that a cultural variant C is either present or not, and assume that cultural transmission is imperfect: When individual a learns from individual b, there is a chance that a will fail to learn. Suppose then that each individual learns with probability p, from a randomly chosen member of a population in which those with C have relative frequency x_t at time t. We'll assume that learning occurs in discrete timesteps, or cultural "generations".

ESELS note that the frequency of those with C will be $p^t x_0$, which decreases to zero unless transmission is perfect.[2] Figure 1 illustrates this process. In each gen-

[1] This concept of a cultural variant does not carry the same connotations as the term "meme", which tends to suggest discreteness, near-perfect replicability, or goal-directedness (cf. Richerson and Boyd 2005; Godfrey-Smith 2009).

[2] I ignore possible effects of random drift in cultural transmission processes throughout this paper.

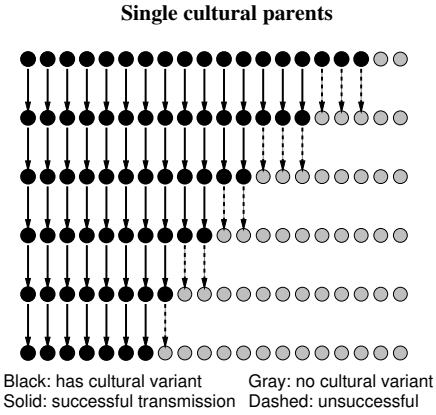


Figure 1: Cultural transmission from randomly chosen "parents" in discrete generations, with time moving from top to bottom. A cultural variant can only be acquired from those who have it (black circles), but transmission (arrows) sometimes fails (dashed arrows).

eration, a fraction p of those with the cultural variant C succeed in transmitting it, so the frequency of C decreases over time. ESELS consider several variations on this model, including (1) models in which multiple learning trials from the same cultural parent are allowed, (2) models in which cultural parents can be chosen because they appear to be successful ("biased transmission"), (3) models in which social learning is combined with individual learning, and (4) models in which possession of a cultural variant confers a fitness advantage that makes the bearer more likely to be available to be chosen as a cultural parent. I won't review all of these models in detail. I note that ESELS showed that that option (1) merely slows down the eventual loss of culture. Option (2) can maintain culture, but only under some parameter values. With option (3), it turns out that under plausible parameter values, culture can only be maintained if individual learning alone could have maintained it. I turn now to option (4).

Suppose that having the cultural variant C provides a fitness benefit. For exam-

ple, perhaps it makes survival more likely, thus allowing an individual more time to influence others. Let $\alpha \geq 1$ be the ratio between the fitness of those with the cultural variant and the fitness of those without it. Under these assumptions, ESELS argue that if the frequency in the current generation is x, the average frequency x' in the next generation is:

$$x' = \frac{p\alpha x}{p\alpha x + (1 - px)} \,. \tag{1}$$

Here $p\alpha x$ is the frequency of successful cultural transmission: αx represents frequency of C among chosen cultural parents, and p is the probability of successful transmission of C. $(1-px)$ is the frequency of transmissions that don't occur, either because the (randomly) chosen parent lacks C, or because the cultural transmission nevertheless failed.

The frequency of C in the population will keep changing until an equilibrium is reached in which $x' = x$. ESELS show that this equilibrium $\hat{x}$ is:

$$\hat{x} = \frac{\alpha p - 1}{p\,(\alpha - 1)} \tag{2}$$

ESELS argue that the equilibrium is nonzero only for parameter values that are rare in nature. For example, if the probability p of successful transmission 0.9—a very high value—the ratio between fitnesses of those with and without C must be greater than about 1.112 in order for culture to be maintained. This is an unusually large fitness advantage.

On the other hand, ESELS show that allowing individuals to learn from two or more cultural parents can easily maintain culture, even without a fitness advantage. The reason is that even if a chosen cultural parent lacks the cultural variant C, or fails to transmit it, there is the the possibility that another randomly chosen cultural parent will successfully transmit C (figure 2). Having a "backup" teacher allows learners to recover, often, from a failure of cultural transmission. In this model, for

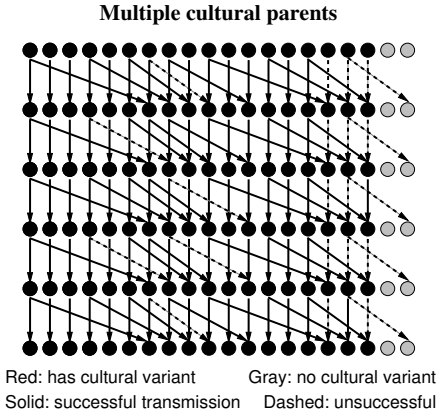


Figure 2: Each diagonal arrow represents cultural transmission from a second cultural parent. See caption for figure 1 for the meaning of other components.

n randomly chosen cultural parents, the frequency x' in the next generation is:[3]

$$x' = 1 - (1 - px)^n \tag{3}$$

ESELS show that there is a nonzero equilibrium if and only if $pn > 1$. On this model, culture will be maintained as long as the probability of transmission is not too low, and the number of cultural parents is high enough. For example, if each learner has two cultural parents, the probability of successful transmission need only be slightly greater than 1/2. Note that this model implies, in effect, that maintenance of culture requires robustness resulting from multiple processes with at least poorly correlated errors (Wimsatt, 2007).

[3]The frequency of successful transmissions is equal to the probability $\mathsf{P}(T_1 \vee \cdots \vee T_n)$ of successful transmission T_i from any of the n cultural parents. Since successful transmission from more than one parent is not ruled out, the probability of at least one successful transmission is equal to $\sum_{i=1}^{n} \mathsf{P}(T_i)$ minus a complex function of probabilities of conjunctions of the T_i's. A routine technique is to simplify such a calculation using De Morgan's law, transforming the above disjunction into $\mathsf{P}(\neg(\neg T_1 \& \cdots \& \neg T_n))$. The frequency is then equal to (3), since the events T_i are independent.

3 Two puzzles

There are two implications of ESELS' argument that may initially seem puzzling. First, many mathematical models of cultural change seem to allow single-parent culture to be maintained. What is the crucial difference in ESELS' model that prevents this? Consider, for example Rogers' (1988) model, which concerns two behaviors, each of which is better adapted to a different environment. Some individual chose a behavior based on individual learning. Others simply copy the behavior of a randomly chosen individual. Rogers shows that at equilibrium, the fitness of social learners is equal to that of individual learners. If we think of the lack of cultural variant C in ESELS' models as a a kind of null cultural variant (which ESELS call N), the loss of C from a population corresponds to fixation of its absence, i.e. fixation of N. However, in Rogers' models, neither of the two cultural variants represented goes to fixation, under a wide range of parameter values. Rogers' model and ESELS' models clearly depend on different assumptions. What are the assumptions in ESELS' model that allow culture to disappear? A closely related question is this: Why not simply switch these labels in ESELS' model? Then it would be inevitable that culture would spread with single-parent transmission! I address these questions in in section 4.

The second implication of ESELS' argument is that it seems to conflict with real-world cases of single-parent transmission. Is it really true, empirically, that culture is rarely maintained through single-parent transmission? To take an example from the cultures of modern industrialized societies, from how many people did you learn long division? How many people taught your teacher? Granted, this isn't much of an argument: Rather than collecting real data, I'm appealing to anecdotal evidence. Nevertheless, there is some reason to wonder whether, or how, the real world might conflict with ESELS' conclusions. I address these questions in section

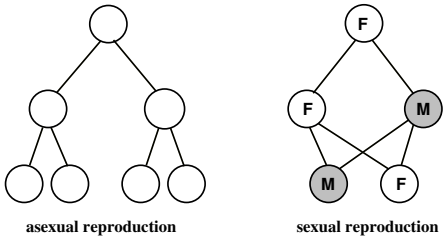


Figure 3: Biological descendant relations in a species limited to two offspring per egg-producing organism.

5.

4 Muller's ratchet

In this section I describe a widely-investigated question concerning the evolution of sexual reproduction. I then describe Muller's ratchet, a class of models that are central to one of several proposed answers to the question. I draw inferences concerning ESELS argument concerning single-parent culture from the close conceptual and mathematical parallels between a recent formulation of Muller's ratchet and one of ESELS' models described above.

Sexually reproducing species evolved from asexual species, and in some species, both kinds of reproduction occur. Even in the case of organisms that simply leave their eggs behind, producing eggs requires significant energy and material resources, which limits the number of offspring that females or asexual organisms can produce. An argument due to Maynard Smith (1978) showed that asexual organisms can have twice as many grand-offspring as sexually reproducing females, with no additional energy expenditure. This is illustrated in figure 3. Suppose, for example, that each egg-producing organism can produce two offspring in a population of asexual reproducers. Each organism will then produce four grand-offspring. If one organism has a new mutation that causes offspring to reproduce sexually,

and produces one male and one female, they must mate to produce offspring, producing only two offspring between them. This argument can be generalized (Maynard Smith, 1978). The upshot is that since sexual reproduction reduces the number of descendants of an organism by half, it must generate an enormous fitness benefit in order for it to have been selected for. There are a number of proposed explanations of the benefit of sexual reproduction that are under active investigation (Otto and Lenormand, 2002). I focus only on Muller's ratchet, a class of models that show that in the absence of sexual reproduction and recombination, deleterious (disadvantageous) mutations will accumulate in a population and eventually cause its extinction.

Muller's ratchet is based on the biologically plausible assumptions that beneficial mutations are rare, that strongly deleterious mutations will quickly be selected out of a population, and that backmutations that undo a mildly deleterious mutation are improbable. Over time, lineages accumulate different numbers of mildly deleterious mutations. Since the evolutionary costs of these mutations are small, lineages with the fewest deleterious mutations will occasionally be lost due to random genetic drift. At that point, all members of the population have more deleterious mutations than those fitter members that were lost, so there is almost no possibility of creating descendants with fewer deleterious mutations. We say that Muller's ratchet has clicked. This process gradually leads to the accumulation of deleterious mutations, and ultimately, to the extinction of the population. Sexual reproduction with recombination provides an escape from the ratchet, though. Recombination combines different segments of two individuals' chromosomes in order to produce an offspring's chromosomes. This allows some offspring to have fewer deleterious mutations than either of their parents, thus preventing the inevitable decline of fitness in the population as a whole that would result from Muller's ratchet.

Most Muller's ratchet models (e.g. Haigh 1978) divide the population into classes of organisms, each with a different number of deleterious mutations. Waxman and Loewe's (2010) "truncated Ratchet" model instead has only two classes: The class of individuals with the fewest deleterious mutations, and a class containing all other individuals. A click of the ratchet is the loss of the fittest class, and a subset of the other class then becomes the new fittest class. Let μ be the per-organism probability of a deleterious mutation; $1 - \mu$ is then the probability that an individual will remain in the fittest class. Let σ be the average fitness cost from deleterious mutations due to being in the less fit class of organisms, and $1 - \sigma$ the fitness of the members of the fitter class. The probability that a deleterious mutation will be undone by a backmutation is so small that it's reasonable to treat it as zero, and we ignore beneficial mutations as well.

Waxman and Loewe then derive the following frequency x' of the fittest class in the next generation, starting from the frequency x in the current generation:

$$x' = \frac{(1-\mu)x}{(1-\sigma) + \sigma(1-\mu)x} \tag{4}$$

(It's not completely straightforward to interpret the components of this formula.) By equating x and x', Waxman and Loewe derive the equilibrium frequency $\hat{x}$ of the fittest class:

$$\hat{x} = \frac{\sigma - \mu}{\sigma(1-\mu)} \tag{5}$$

I want to compare ESELS' model with selection and Waxman and Loewe's model. Both sets of models have a parameter representing the probability of error-free transmission. In ESELS' model, this is p, the probability of transmitting the cultural variant C without error. In Waxman and Loewe, $1 - \mu$ is the probability of reproduction without a deleterious mutation. Both sets of models also have a parameter representing the fitness difference between two states. ESELS' α is the ratio between the fitnesses of a cultural variant C and its absence, N. Waxman and

Loewe instead represent the relationship between fitnesses of the fittest and less fit classes with a difference parameter σ. We can equate the two sets of parameters, with ESELS' parameters on the left, and Waxman and Loewe's on the right:

ESELS		Waxman and Loewe
p	$=$	$1-\mu$
$\alpha \times$ (fitness of N)	$=$	1
(fitness of N)	$=$	$1-\sigma$

When we equate the parameters, it turns out that Waxman and Loewe's truncated ratchet is mathematically equivalent to ESELS' single-parent model with fitness: Equation (1) is equivalent to equation (4), and equation (2) is equivalent to equation (5). This close relationship between ESELS' model and Waxman and Loewe's truncated ratchet shows that failing to learn from a single cultural parent is so closely analogous to acquiring a deleterious mutation in an asexual species, that both relationships can be modeled in the same way. Similarly, in either model, the loss of beneficial characteristics can be avoided by allowing information to be transmitted from (at least) two parents.

Importantly, the emphasis in Muller's ratchet models on relations between certain probabilities highlights similar relations in the ESELS models. As in other Muller's ratchet models, Waxman and Loewe's model makes the probability of a deleterious mutation (small, but positive) and no mutation (large) significantly different. Importantly, Muller's ratchet models set the probability of undoing a deleterious mutation equal to zero. Analogously, in ESELS' single-parent model with with fitness, transmission of a beneficial trait C has a significant probability of failure, and there is no chance of undoing the loss of C from a lineage. This is what creates an inevitable loss of culture, just as Muller's ratchet creates an inevitable loss of fitter variants. By comparison, Rogers' (1988) model gives equal probability

to learning either of two cultural variants. Models of that kind are appropriate for cultural variants that can easily replace each other in a given cultural context. ES-ELS' models, on the other hand, seem most appropriate for cultural variants that are difficult to learn in the first place, and that are readily lost without any replacement.

5 Coherence

I noted above that it seems somewhat plausible that there are cases of single-parent transmission. In this section I suggest that probabilistic coherence measures inspired by C.I. Lewis's (1946) concept of congruence can, with a slight reinterpretation, be used to understand how a kind of single-parent cultural transmission—or at least the illusion of single-parent cultural transmission—can maintain culture indefinitely.

C.I. Lewis defined his coherence measure, known as *congruence*, as follows:

> A set of statements, or a set of supposed facts asserted, will be said to be congruent if and only if they are so related that the antecedent probability of any one of them will be increased if the remainder of the set can be assumed as given premises. (Lewis, 1946, p. 338)

Two propositions C_1 and C_2 are thus congruent iff:

$$\mathsf{P}(C_1|C_2) > \mathsf{P}(C_1)$$

$$\mathsf{P}(C_2|C_1) > \mathsf{P}(C_2)\ .$$

For two propositions, this relationship can also be captured by requiring that Shogenji's (1999) coherence measure

$$\frac{\mathsf{P}(C_1\&C_2)}{\mathsf{P}(C_1)\mathsf{P}(C_2)}$$

be greater than 1.

Lewis's congruence does not fit all intuitions about the role of coherence in justification (Olsson, 2005). Olsson (2005) has shown, for coherence measures such as Shogenji's, that greater coherence does not consistently imply a higher probability of truth of the propositions considered. However, my concern below will not be with truth-conduciveness. I'll treat the probabilities above as probabilities of believing one proposition given believing the other, or more generally, as probabilities of acquiring one cultural variant given the other. This is in effect translates a justificatory relation into a cognitive or behavioral relation.

In real human cultural transmission, we don't only learn isolated bits of information. Research suggests that we remember and can use what we learn more effectively if it is systematically related to other things we learn (Bransford and National Research Council, 2000). Roughly, it helps if different things we learn are related and mutually supporting.[4] I'll suggest that we can capture some aspects of this property of human learning with Lewis's congruence notion, and with various extensions of it. In what follows, I'll focus on the simple case of two cultural variants C_1 and C_2 that influence each others' adoption.

First note that ESELS's arguments can be applied simultaneously to two cultural variants C_1 and C_2: If either cultural variant is transmitted only by single cultural parents, it will eventually be lost from the population. This is true whether the two variants are both learned from the same parent, or from distinct randomly chosen parents. Research on learning mentioned above, however, raises the possibility that some cultural transmission of coherent beliefs might help prevent the loss of culture. If C_1 and C_2 are congruent, so that each raises the probability of believing the other, could this prevent the loss of culture? It appears that it can only slow

[4]Proviso: The research summarized in (Bransford and National Research Council, 2000) seems to focus only on formal schooling in industrialized societies. Henrich et al. (2010) argue that many experimental results from industrialized populations do not generalize to all humans.

down the loss of culture, if both beliefs must be acquired from the same cultural parent. I'll explain why.

Suppose that a learner can acquire all or some of the cultural variants possessed by a randomly chosen cultural parent, where the probability of acquiring any one of the cultural variants is the same, p. Suppose also that it's possible for an individual to infer (etc.—see below) a missing cultural variant C_i if one has the other cultural variant. We can capture the probability of such an inference by:

$$\mathsf{P}(C_1|C_2) = \mathsf{P}(C_2|C_1) = r > 0\ . \tag{6}$$

This is like congruence, or Shogenji's coherence being greater than 1, since we're assuming that the probability of spontaneously acquiring C_1, i.e. $\mathsf{P}(C_1|\neg C_2)$, is zero (note $\mathsf{P}(C_1|C_2) > P(C_1)$ iff $\mathsf{P}(C_1|C_2) > P(C_1|\neg C_2)$).[5] . These conditional probabilities in effect link C_1 and C_2, so that if a person fails to acquire one of them, but acquires the other, she'll be able to acquire the first anyway, with probability r. We can call r a "generalized inference probability", since if C_1 and C_2 are beliefs that would cause each other to be held due to rational inference, r is the probability of inferring one from the other, even though there may be other reasons that adopting one cultural variant would cause the other to be adopted.

Let's focus on the best case for using such internal inference processes to maintain culture: Let r be near 1. Then most individuals with either cultural variant will have the other. Those who have only one of C_1 and C_2 will nevertheless be able to produce cultural offspring who have both cultural variants. However, with large r, C_1 and C_2 are functioning almost as a unit. The process is roughly the same as what would happen in ESELS' single-parent model if we increased the transmission probability p: It would take longer for culture to disappear, but it would still do so, eventually.

[5]Elsewhere I spell out this point in more detail. Feel to contact the author for a longer exposition.

Suppose, however, that equation (6) holds, but that the cultural parent for each of the two cultural variants is chosen independently for each cultural "child". Different cultural variants are acquired from different cultural parents. Thus the probability of acquiring one cultural variant (e.g. C_1) given the other (C_2) is r. And assume that the parent-to-child transmission probability for each variant is p. Then the probability of acquiring C_1 is similar to probability of acquiring C from either of two cultural parents in equation (3). Let x be the relative frequency of C_1 in the population, and y the relative frequency of C_2. Then the frequencies x' of C_1 and y' of C_2 in the next generation are:

$$x' = 1 - (1 - px)(1 - rpy) = px + rpy - rp^2xy \quad (7)$$

$$y' = 1 - (1 - py)(1 - rpx) = py + rpx - rp^2xy \quad (8)$$

The first equation, for example, is based on the following reasoning (cf. note 3). One can acquire C_1 directly from the cultural parent chosen to transmit C_1, with probability px, or fail to do so with probability $1 - px$. Similarly, one can acquire C_1 with probability r from C_2 if C_2 was acquired—which happens with probability py. So failing to succeed by this path is $1 - rpx$. The probability of successful transmission of C_1 is then is the probability of failing to fail to acquire C_1 by one of the two probabilistically independent paths.

The dynamics of this model are not identical to those of the simple two-parent transmission model characterized by equation (3). However, we can simplify the new model, because iterating equations (7) and (8) quickly causes C_1 and C_2 to have the same frequency. To see this note that

$$|x' - y'| = |px + rpy - py - rpx| = p(1 - r)|x - y| .$$

But $p(1 - r)$ must lie between 0 and 1, so the difference between the frequency x of C_1 and the frequency y of C_2 shrinks in every generation.[6] Thus for the long

[6] In simulations with a variety of values of p and r, x and y come together within 10 or 20 gener-

term effects of cultural transmission of C_1 and C_2 when the inference probability r is the same in both directions, we can ignore the difference between their initial frequencies, and model the change in either frequency as:

$$x' = 1 - (1 - px)(1 - rpx) \tag{9}$$

When r is high, this equation shares with (3) the property that when one fails to learn a cultural variant from a single chosen cultural parent, another parent is relatively likely to transmit that variant. In this case, however, the transmission from the second parent is indirect, via the other cultural variant C_2 (cf. figure 2).

The frequency of C_1 (or C_2) is at equilibrium when $x' - x = 0$, i.e. when

$$0 = (px + rpx - rp^2x^2) - x \;=\; x([rp + p - 1] - rp^2x) \tag{10}$$

The right hand side is equal to zero either when $x = 0$,[7] or when $(rp+p-1)-rp^2x = 0$, i.e. when x has the value:

$$\hat{x} = \frac{rp + p - 1}{rp^2} \; . \tag{11}$$

For example, this equilibrium frequency is approximately 0.9 when $p = r = 0.83$. This equation also shows that the equilibrium is greater than zero iff $rp + p > 1$, i.e. iff

$$r > \frac{1 - p}{p} \; . \tag{12}$$

Thus there is a nonzero equilibrium whenever the internal inference probability is greater than the ratio between the probabilities of direct transmission failure and success, which holds whenever p and r are not too small. When p and r are equal, they must be greater than about 0.62 for the cultural variants to stabilize at

ations, even with initial values $x = 0.01$ and $y = 0.99$.

[7]When $x = 0$, equation (9) is misleading. If (7) and (8) are allowed to iterate, it's possible for one cultural variant, say C_2, to begin with frequency $y = 0$ and still reach a non-zero equilibrium.

a nonzero value.[8]

I suggest, then, the fact that some cultural variants—perhaps long division—seem to be maintained by transmission from single cultural parents, may be due to the fact that these variants are supported by transmission of other variants from other cultural parents. Long division, for example, is not learned in a vacuum. A variety of closely related mathematical concepts are usually learned first, and subsequent use in other contexts provides additional support for it. Thus it may be that students are able to infer, or at least be reminded of, missing steps in long division when forgotten because of what they learned from multiple cultural parents.

The preceding model considered only two related cultural variants, but real learning and real culture surely involve more complex relations of support between variants learned. Lewis's (1946) concept of congruence, perhaps formalized as Angere's (2008) C_E, allows for probabilistic support involving more propositions. Shogenji's (1999) measure of coherence does as well, but in a different way. Schupbach (2011) extends Shogenji's measure to make it sensitive to additional relations of probabilistic support. Fitelson's (2003; 2004) coherence measure reflects more relations of probabilistic support between conjunctions of propositions in a given set. It may be that one of these coherence measures, or a related one, though not designed to capture the degree to which a set of cultural variants allow restoration of one from others, will be useful for characterizing such a property.

Although classic treatments of the role of coherence in epistemology discuss it in probabilistic terms (Lewis, 1946; BonJour, 1985), the paradigmatic relations underlying the probabilities were usually thought to be, or to be similar to, logical or explanatory relations (cf. also Lehrer 2000). This makes sense given that coherence

[8]It can be shown that when C_1 and C_2 have different inference and direct transmission probabilities, a nonzero equilibrium exists iff $rs > (1-p)(1-q)/pq$, where r and s are inference probabilities and p and q are corresponding transmission probabilities (proof available upon request).

was intended to provide justification for beliefs. In giving ideas from epistemological models of coherence in a role in cultural transmission, we have to allow a broader basis for the relevant probabilities, though. When a person comes to adopt cultural variant C_1 because they have previously adopted variant C_2, this could be because both cultural variants are beliefs, and they have noticed that C_2 helps to justify C_1. This justificatory relationship might be mediated by a great deal of cultural background belief, however. However, other sorts of relationships between beliefs may provide the basis for the probabilistic relationship between acquisition of cultural variants. It may be that given the cultural background that the person has already adopted, there is some nonlogical resonance felt between C_1 and C_2.

6 Conclusion

Enquist et al. (2010) argued that culture will usually disappear unless individuals learn from multiple "cultural parents". The conceptual and mathematical analogy of one of ESELS' models to Waxman and Loewe's truncated Muller's ratchet model clarified that ESELS' models make the assumptions, unusual among cultural transmission models, that transmission probabilities very assymetric, and that once a person loses a cultural variant, it can never come back among cultural "descendants". This crucial assumption makes ESELS' models relevant to cultural variants that are difficult to learn socially—so that they may not be learned at all—and difficult to learn individually. Modeling relations between cultural variants using ideas from probabilistic measures of coherence in epistemology provides a way of modeling the influence of one cultural variant on another. This allows the learning of C_2 from single parents to help maintain C_1 in the population, even if C_1 is itself only learned from single parents. Coherence in this sense captures what might be called "inferential robustness": the ability to infer or otherwise learn cultural

variants through multiple, redundant paths (Wimsatt, 2007). This way of thinking about coherence may also be relevant to epistemology.

References

Angere, Staffan (2008). Coherence as a heuristic. *Mind* 117(465):1–26.
URL http://mind.oxfordjournals.org/content/117/465/1.abstract

BonJour, Laurence (1985). *The Structure of Empirical Knowledge*. Harvard.

Bransford, John and National Research Council, (U.S.) (2000). *How People Learn : Brain, Mind, Experience, and School*. National Academy Press.

Enquist, Magnus; Strimling, Pontus; Eriksson, Kimmo; Laland, Kevin; and Sjostrand, Jonas (2010). One cultural parent makes no culture. *Animal Behaviour* 79:1353–1362.

Fitelson, Branden (2003). A probabilistic theory of coherence. *Analysis* 63:194–199.

Fitelson, Branden (2004). Two technical corrections to my coherence measure. (From the author's website, fitelson.org. Downloaded March 1, 2014.).
URL http://fitelson.org/coherence2.pdf

Godfrey-Smith, Peter (2009). *Darwinian Populations and Natural Selection*. Oxford University Press, Oxford, UK.

Haigh, John (1978). The accumulation of deleterious genes in a population—Muller's ratchet. *Theoretical Population Biology* 14:251–267.

Henrich, Joseph; Heine, Steven J.; and Norenzayan, Ara (2010). The weirdest people in the world? *Behavioral and Brain Sciences* 33:61–83.

Lehrer, Keith (2000). *Theory of Knowledge*. Westview, 2nd ed.

Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. Open Court Publishing.

Maynard Smith, John (1978). *The Evolution of Sex*. Cambridge University Press.

Olsson, Erik J. (2005). *Against Coherence: Truth, Probability, and Justification*. Oxford University Press, Oxford, UK.

Otto, Sarah P. and Lenormand, Thomas (2002). Resolving the paradox of sex and recombination. *Nat Rev Genet* 3(4):252–261.

Richerson, Peter J. and Boyd, Robert (2005). *Not By Genes Alone*. Oxford University Press, Oxford, UK.

Rogers, Alan R. (1988). Does biology constrain culture? *American Anthropologist* 90(4):819–831.

Schupbach, Jonah N. (2011). New hope for Shogenji's coherence measure. *British Journal for the Philosophy of Science* 62:125–142.

Shogenji, Tomoji (1999). Is coherence truth conducive? *Analysis* 59.4:338–345.

Waxman, David and Loewe, Laurence (2010). A stochastic model for a single click of Muller's ratchet. *Journal of Theoretical Biology* 264:1120–1132.

Wimsatt, William C. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Going against the Grain: Functionalism and Generalization in Cognitive Science

MIKIO AKAGI

Abstract: Functionalism is widely regarded as the central doctrine in the philosophy of cognitive science, and is invoked by philosophers of cognitive science to settle disputes over methodology and other puzzles. I describe a recent dispute over extended cognition in which many commentators appeal to functionalism. I then raise an objection to functionalism as it figures in this dispute, targeting the assumption that generality and abstraction are tightly correlated. Finally, I argue that the new mechanist framework offers more realistic resources for understanding cognitive science, and hence is a better source of appeal for resolving disagreement in philosophy of science.

Word count: 4,985 words including abstract, headings, footnotes, and references.

1 Introduction. Functionalism is the doctrine that mental or cognitive states are functional states, whose identity conditions are articulable in terms of their characteristic inputs, outputs, and relations to other intermediate states. Functionalism was established as a central doctrine in the philosophy of cognitive science in the 1960s (Putnam 1967a, b, Fodor 1968), and though it has become less central to much contemporary discussion (Chemero & Silberstein 2008) it retains the notoriety of an orthodoxy in philosophy of mind (Buechner 2011) and in contemporary philosophy of cognitive science (Eliasmith 2002, Clark 2008, Sprevak 2009, Chalmers 2011). This remains true even though functionalism has been an embattled doctrine for decades (Block & Fodor 1972, Block 1980, Shagrir 2005, Godfrey-Smith 2008), has proliferated versions and variations (Levin 2013, Maley & Piccinini MS), and even though the canonical argument for functionalism—the argument from multiple realizability—has been subjected to a variety of criticisms (Batitsky 1998, Bechtel & Mundale 1999). This is all, importantly, to say nothing of other views that happen, unhappily, to be called "functionalism" in biology or in pre-behaviorist psychology (Cummins 1975, Sober 1985, Chemero 2009) but which have different intellectual lineages. My discussion concerns only Putnam's machine functionalism and derivative views. The persistence of functionalism is hardly a special case. It is the fate of many "received views," such as the belief-desire model of intentional action or the deductive-nomological model of explanation, to remain central to a literature despite decades of convincing criticism so long as there is no sufficiently dominant successor. The new mechanist view of explanation (Machamer, Darden, & Craver 2000, Bechtel & Abrahamsen 2005, Craver 2007) has

recently achieved this status in the philosophies of the biological sciences, supplanting the deductive-nomological model and other law-subsumption models as a received view of explanation in those sciences. This is not to say that new mechanism is correct or uncontroversial, only that it has replaced other models of explanation as the primary target of interpretation and criticism.

Bearing in mind the tenacity of received views, my aim in this paper is not to simply poke more holes in the sinking ship of functionalism. Rather, I aim to promote an alternative vessel for philosophers of mind and cognitive science to pilot through choppy waters. To this end I will raise a "grain" objection to functionalism, based on the relationship between generalization and "fineness-of-grain." This objection is not a knock-down argument against all varieties or uses of functionalism. However, it needn't be, since functionalism is a sinking ship, and since my objection does apply to versions of functionalism that figure in notorious, recent disputes in the philosophy of cognitive science. I shall take as my example the controversy over extended cognition, especially its recent high-profile epicycle concerning the relation between extended cognition and functionalism (Rupert 2004, Clark 2008, Sprevak 2009). In the last section I will argue that new mechanism provides better resources for understanding variation between models in cognitive science, and for understanding the practice of generalization. In particular, mechanism is not vulnerable to the grain objection. I do not claim, of course, that mechanism is free from criticism or worries or that I have made clear what was once obscure. My aim, rather, is to motivate a change of focus in discussions of cognitive science from functionalism to mechanism.

2 Functionalism and Extended Cognition. Andy Clark and David Chalmers (1998) notoriously claim that cognition (like meaning) ain't all in the head. They argue that in certain cases the use of external props in some activities—a computer processor while playing some video games, one's notebook in carrying out one's plans for the day, perhaps one's partner in remembering past events—is such that those props should be considered parts of one's own cognitive economy, similarly to parts of one's brain. This claim has become known as the hypothesis of extended cognition (HEC). The most famous example concerns Otto, an older gentleman with a bad memory who uses a notebook to help him remember facts and plans.[1] In their argument, Clark and Chalmers appeal to what has become known as the "parity principle," which states that

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of a cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. (Clark & Chalmers 1998, 8)

[1] The Otto example is originally an illustration not of extended cognition but of the extended mind, which is a distinct claim. Although this distinction is essential for charitably evaluating Clark and Chalmers' arguments, it is almost always ignored in the critical literature (even by Clark). Since I am not evaluating HEC here, I will ignore the distinction for ease of exposition.

One way to interpret this principle is as a corollary of functionalism: cognitive states are individuated by their functional relations (to inputs, outputs, and each other), and it is immaterial whether their realizers are located inside the brain or outside the body.[2] Thus, activities should count as cognitive processes if those body-external processes exhibit the same functional relationships (to inputs, outputs, and cognitive states) as other processes that we already happily consider cognitive processes. Of course, understood this way the parity principle only justifies a commitment to extended cognition if the functional relations are specified so that body-external activities and props do satisfy those specifications, and many cognitive and psychological processes can be specified in a variety of ways. Robert Rupert argues that Otto's notebook in the famous example cannot serve as a memory in part because it fails to satisfy the most fruitful functional description of human memory. Cognitive psychologists have documented many features of human memory—for example susceptibility to interference effects, generation effects, and conformity to the Rescorla-Wagner law (see Rupert 2004, 413–419). Since Otto's external "memory" does not exhibit these effects the parity principle does not license the attribution to Otto of extended cognitive processes (Rupert 2004). Fred Adams and Ken Aizawa (2001) argue for the same conclusion because Otto's use of his notebook must be described via *inter alia* relations to perceptual and motor intermediaries (he flips through the notebook, reads it, &c.), whereas canonical examples of internal memory are not related to perceptual and motor activities in the same way.

Mark Sprevak calls these objections the RAA (for Rupert, Adams, and Aizawa) objections. Sprevak suggests that "All varieties of functionalism contain a parameter that controls how finely or coarsely functional roles should be specified (how much should be abstracted and ignored)" (Sprevak 2009, 510). He observes that RAA trade on *fine-grained* differences between Otto's use of his notebook and canonical examples of memory. A coarse-grained functional description of memory might simply describe the relations between past perceptions and actions and future behaviour, but not describe memory as e.g. exhibiting interference or generation effects, or as obeying the Rescorla-Wagner law. Fine-grained functional descriptions may specify these relations, but are objectionable because they conflict with the common intuition that there could be Martians who have cognitive processes but whose cognitive architecture is distinctly different from ours. Such Martians, unlike us, may not exhibit interference or generation effects, and may even store information by manipulating ink-marks on paper inside of their brains, and retrieve it by reading the marks back with photosensitive organs. The "Martian intuition" is that while this is an alien form of memory, it is memory nonetheless. Since such Martians have memory, and their memory may have the same fine-grained functional description as Otto's use of his notebook, the parity principle demands that we consider Otto's a case of extended cognition. Thus Sprevak argues that functionalism implies HEC.

Unfortunately for the defenders of HEC, however, Sprevak argues that coarse-grained functional descriptions are no more acceptable, for the parity principle is less restrictive than Clark and Chalmers anticipate. Since we can imagine far-fetched Martian minds, the parity principle

[2] This is accepted by most of Clark and Chalmers' critics (Adams & Aizawa 2001, Rupert 2004, Sprevak 2009), and is almost certainly not the best interpretation of the parity principle.

licenses a radical form of HEC. For example, we might imagine Martian minds that are embedded with factual information that must be retrieved with effort, so that this process has functional parity with the activity of looking up information in a library. Such possibilities seem to license radical cases of extended cognition: that contents of volumes in a library are beliefs of any person in the library, or that being in possession of a graphing calculator gives one a knowledge of integral calculus (517–518). These consequences, Sprevak argues, are absurd, and justify a *reductio* of radical HEC and, since it entails radical HEC, of functionalism.

3 Going against the Grain. The dispute over HEC is not seen by its partisans as an idle philosophical discussion, but as a battle for the soul of cognitive science. If HEC is true, it is claimed, it has dramatic consequences for the way cognitive scientists conduct their research. Hence, both defenses and criticisms of HEC draw on empirical results and claims about theory-choice in science (e.g. Clark & Chalmers 1998, Rupert 2004, Adams & Aizawa 2008, Clark 2008, Rowlands 2010). The fact that so many of the arguments concerning HEC trade on interpretations of functionalism reveals the belief of many that functionalism provides a suitable framework for understanding cognitive scientific models. Disagreements about HEC force a discussion of what precisely the laws of cognitive science are—both what their proprieties are with respect to generalization, and what phenomena should be investigated and accommodated in order to construct those laws. In Putnam-style functionalism, functional descriptions (i.e. via Ramsey sentences, cf. Lewis 1972) operate as laws characterizing mental states. Putnam's proposal aspires explicitly to generalization over diverse kinds of system—e.g. about pain in primates and also in cephalopods. The arguments that arise in connection with the RAA objections to HEC concern whether descriptions of e.g. memory generalize over head-internal vehicles and extended vehicles (like Otto's notebook-use).

Like Clark and Sprevak, Hilary Putnam is wary of psychological chauvinism (human-specificity). A type-physicalist account of pain (Place 1956, Smart 1959), like the simplistic conjecture that pain is the activation of C-fibers, denies without motivation that animals that lack C-fibers have pain-states. In Putnam's canonical argument for functionalism, functionalism achieves *generality* by proposing *abstracted* descriptions that omit physiological and other details. Sprevak's grain parameter makes this maneuver more explicit by proposing a continuum of descriptions that are increasingly abstract, in the sense of omitting detail, and therefore increasingly general. (Some may object to the use of the word "abstract" as the complement of "detailed," rather than of "concrete." However, Sprevak uses the word this way and there is ample precedent for his doing so (Levy & Bechtel 2013).) Consider a toy functional description of pain: pain is caused by tissue damage, and causes stress, increased metabolic activity, and evasion of the damaging stimulus. This description denotes processes in a variety of complex organisms, including cephalopods (which lack C-fibers). Elements can be added to this description to make it more fine-grained, and to denote processes in progressively more restricted classes of organisms. For example, if pain also tends to cause excited

vocalization, then creatures like cephalopods which do not vocalize will not satisfy this more fine-grained description of pain.

However, Sprevak's grain parameter is not an effective way of capturing variation between cognitive models. In the space of models that cognitive scientists actually produce, generality-specificity and abstraction-detail are independent dimensions of variation. By way of example, I shall mention two cognitive models in which generality and abstraction are dissociated. The first, the motor theory of speech perception (Liberman et al. 1967, Liberman & Mattingly 1985), is quite abstract but highly specific to humans. The motor theory claims that "perceiving speech is perceiving gestures," and more specifically that the recognition of phonemes and words in natural language is mediated by processing in the motor system, namely motor processing that also governs the articulation of speech in the vocal tract. There are animals other than humans that can identify phonemes and words—dogs commonly learn to recognize some words, and chinchillas have been trained to distinguish natural language phonemes (Kuhl & Miller 1978)—but since they do not have the relevant vocal capacities they most likely exhibit this capacity exclusively by recognizing auditory patterns, whereas humans do not. Nevertheless, the motor theory is quite abstract—it specifies that speech perception depends on structures that govern vocalization. While there are more detailed claims about how this dependency manifests in humans (e.g. McGurk & MacDonald 1976), all of them are consistent with the motor theory.

On the other hand, feature-detector models of vision (e.g. Barlow 1953, Hubel & Wiesel 1962) are detailed, but general. Even normalization-based models of feature-detection in particular (Heeger 1992, Carandini & Heeger 1994), which are described by Mazviita Chirimuuta (2014), are quite general. These models describe sensitivity to contrasts, edges, &c. in early stages of visual processing, and unify evidence about the receptive fields of individual neurons as well as computational models of their response dynamics. On Heeger's normalization model, neurons in visual cortex respond linearly to excitatory input from the lateral geniculate nucleus, but inhibit each other "laterally" according to an equation. The terms of the equation stand for properties and activities of individual neurons and populations of neurons. This model can be integrated into conjectures about the gross architecture of visual cognition (Marr 1982), and features in the "standard model" of primary visual cortex (Rust & Movshon 2005). However, even without supplementation with other models of visual processing the normalization model makes quantitative predictions about neuronal activity and has a well-specified physiological interpretation. Nevertheless, despite the level of detail in contemporary feature-detector models, they do not apply only to humans. Early evidence for normalization was gathered largely from cats and frogs, and the models may generalize to all vertebrate vision.

I am sympathetic to Sprevak's conclusion that functionalism is false, however functionalism is in worse shape than he acknowledges. His argument presupposes that we can manipulate abstraction from detail like the mesh of a sieve to sift the chauvinistic cognitive models from the liberal models. However, the motor theory of speech perception and feature-detector models of visual processing illustrate the double dissociation between abstraction from detail and generality

over diverse kinds of cognitive systems. If the grain parameter is supposed to track degrees of abstraction from detail, then it fails to simultaneously track generalizability in cognitive models. If it is meant to track both, it fails to accurately capture the variation in cognitive models. Either way it incorporates false presuppositions about the character of the variety in cognitive models.

4 Generality without Laws. The problems with functionalism that are made explicit in the "grain" objection are inherited from the covering-law view of explanation and generalization that was popular throughout the twentieth century. On that conception, generalization is achieved by subsuming many phenomena under a common description (expressing a "covering law"). However, the covering-law view has in recent years been supplanted by the new mechanist view of explanation, at least in the biological sciences. The mechanists hold that many scientific explanations, including a preponderance of explanations in the biological sciences, are achieved by specifying models of mechanisms. The extension of the mechanist view to cognitive science requires the suppression of certain controversial assumptions developed for biological contexts (especially certain assumptions of Craver 2007, see Weiskopf 2011, Chirimuuta 2014), but not all mechanists make these assumptions (cf. Machamer, Darden, & Craver 2000, Bechtel 2008).

Let us suppose that the primary explananda of cognitive science are intelligent behaviors or cognitive capacities. Intelligent behavior is behavior that is sensitive to the circumstances of an organism and that can be rationalized by its relation to a goal of the organism; cognitive capacities are those that are exhibited in intelligent behavior. A cognitive mechanism, then, is a structure of component entities and component operations that are organized such that they produce intelligent behavior (adapted from Machamer, Darden, & Craver 2000, Bechtel & Abrahamsen 2005, Craver 2007). The entities that figure in cognitive mechanisms are things like representations, modules, brain areas, populations of neurons, or idealized "neurons" in artificial neural networks. Characteristic operations in cognitive mechanisms are processing operations on or between those entities: transformations of representations, computational interactions between modules and brain areas, activation and inhibition of neuron populations, and interactions between artificial neurons as specified by connection weights. The organization of these entities and operations into mechanisms is usually represented by graphs, but can be specified more or less completely by groups of equations or descriptions of relations between components. Cognitive models are models of how mechanisms produce cognitive capacities (possibly or actually), and functional roles can be assigned to components of the models according to how those components contribute to the mechanism's production of that capacity (roughly as described in Cummins 1975). Cummins-style functional roles, however, are not functional descriptions; they describe a component's contribution to a capacity rather than conferring identity conditions in virtue of relations to input, outputs and intermediate states, and are thus independent from Putnam-style functionalism (see Craver 2001 for a discussion of Cummins-functions and neo-mechanism).

If the mechanist framework is to overcome functionalism's difficulties with generalization, it must provide an alternative to the covering-law framework, or even a covering-model framework

(Bechtel & Abrahamsen 2005, Craver 2007, 66–70). After all, it is now widely believed that the biological and social sciences have no true laws. William Bechtel and Adele Abrahamsen (2005) suggest that mechanistic explanations are generalizable not because the target systems are *identical* in the relevant respects, but because they are *similar*:

> The need to invoke similarity relations to generalize mechanistic explanations seems to be a limitation of the mechanistic account. But in fact it may be the mechanistic account that provides a better characterization of how explanations are generalized in many sciences. Laws are generalized by being universally quantified and their domain of applicability is specified by the conditions in their antecedents. On this account, no instance better exemplifies the law than any other. But in actual investigations of mechanisms, scientists often focus on a specific exemplar when first developing their accounts. (Bechtel & Abrahamsen 2005, 438)

The claim that generalization is based on similarity to exemplars is less satisfying than the picture of subsumption under a covering law. Bechtel and Abrahamsen's claims do little to constrain the practice of licit generalization, and their observation that scientists "seem to have an intuitive sense" of how to generalize is distinctly unsatisfying (ibid). However, given the lack of universal or exceptionless laws in the biological sciences, a more complicated conception of generality is needed. The need to be more specific about "similarity"-based generalization is not a drawback of the mechanist framework, but a demand for further research by philosophers of science.

The mechanist framework offers richer resources than functionalism for constrained similarity-comparisons. First, mechanism models are more structured than functional descriptions. Functional descriptions might be structured according to independent predicates or conjuncts inside the scope of the quantifier in a Ramsey sentence. In comparing two mechanisms, one can appeal to similarities and differences between the sets of entities, of operations, their properties, or in their organization. Importantly, the result of such comparisons is not a judgment that mechanisms described by different models are simply the same or different, but that they are similar in certain respects and dissimilar in others. Frequently, a model may apply but with modifications, with the consequence that insights are gained both for the new and for the original target systems. For example, the two visual streams hypothesis (Milner & Goodale 2006) was developed for primate visual systems, primarily with data from humans and macaques, but comparisons of primates and other organisms such as frogs enrich the model (see e.g. Goodale & Humphrey 1998, 183–185) and provide a framework around which similar models can be developed for most vertebrates (Jeannerod & Jacob 2005, 301). Generalization here is achieved through comparisons to exemplars with acknowledgement of differences, not subsumption under a common description. This is mechanistic generalization by, if you like, functional similarity, but not functional identity in Putnam's sense. Since similarity-based generalization like this does not presuppose that generality and abstraction from detail are correlated. Thus the motor theory of speech perception and the normalization model of visual feature-detection are not anomalies in the mechanist framework.

It might be possible to provide similarity-based generalizations of functionally-individuated kinds, but such a strategy is not pursued by those who appeal to functionalism in order to settle other questions in philosophy of cognitive science. For example, the strategy is not pursued by Adams and Aizawa, Rupert, or Sprevak in their criticism of HEC, who instead seek categorical descriptions of mental or cognitive processes. Clark and Chalmers appeal to a relatively abstract specification of memory to argue that Otto's notebook functions as a part of his memory. RAA appeal to relatively detailed specifications of memory to argue that he does not. An ecumenically-minded theorist might suggest that alternative specifications—some detailed and some abstract—delineate various dimensions of similarity and difference between paradigmatic memory and Otto's notebook-augmented memory. However, such a proposal must specify how membership is decided for the set of admissible descriptions for a term. The main products of cognitive scientific research (apart from philosophical research) are models, not functional descriptions. The mechanist framework provides a more natural resource for appeal in philosophy of cognitive science than an unarticulated successor to functionalism. In general the place for functionalism as a resource for appeal in philosophy must be reevaluated.

5 Conclusion. My intention in this paper was to show that the assumptions of functionalism are inappropriate for thinking clearly about cognitive science. To this end I described some discussion of the RAA objections to HEC, and claimed that Sprevak's "grain parameter" makes explicit an assumption that features in the motivating arguments for functionalism: that abstraction from detail and generality are correlated features of cognitive models. This assumption is false, so functionalism is an inappropriate framework for characterizing cognitive models and for settling disputes about cognitive science, like the dispute over HEC, that turn on generality. I suggested, following a suggestion by Bechtel and Abrahamsen, that where the functionalist framework hides the complexity in cognitive scientists' practice of generalization, the mechanist literature provides a more fruitful framework for exploring that complexity. I have not argued that mechanism settles whether HEC is true or false. However, if disagreements about HEC are to be a battle for the soul of cognitive science, the proper battleground is over what kinds of mechanisms are cognitive ones, not over functionalist descriptions of mental states (cf. Walter 2010). The mechanist framework does not provide us with resources for determining the identity conditions of cognitive phenomena like belief and memory, as the functionalist framework does. However, cognitive scientists do not take conformity to their models as a criterion of exhibiting a phenomenon. For example, psychologists do not claim that exhibiting interference effects is a necessary condition on memory. That a system does not exhibit interference effects implies that memory models that do exhibit such effects must be modified in order to be generalized to that target system, not that the target system lacks genuine memory. It is therefore peculiarly contentious for philosophers to appeal to these models in order to settle the identity conditions for cognitive phenomena under the guise of being scientific. The contentious nature of this form of argument is no doubt obscured by the common belief that functionalism is an orthodoxy of cognitive science.

REFERENCES

Adams, Fred, and Ken Aizawa. 2001. "The Bounds of Cognition." *Philosophical Psychology* 14: 43–64.

---. 2008. *The Bounds of Cognition*. Malden, MA: Blackwell.

Barlow, Horace D. 1953. "Summation and Inhibition in the Frog's Retina." *The Journal of Physiology* 119: 69–88.

Batitsky, Vadim. 1998. "A Formal Rebuttal of the Central Argument for Functionalism." *Erkenntnis* 49: 201–220.

Bechtel, William. 2008. "Mechanisms in Cognitive Psychology: What Are the Operations?" *Philosophy of Science* 75: 983–994.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Bechtel, William, and Jennifer Mundale. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66: 175–207.

Block, Ned. 1980. "Troubles with Functionalism." In Ned Block (ed.), *Readings in Philosophy of Psychology*. Cambridge, MA: Harvard University Press, pp. 171–184.

Block, Ned, and Jerry A. Fodor. 1972. "What Psychological States Are Not." *Philosophical Review* 81: 159–181.

Buechner, Jeff. 2011. "Not Even Computing Machines Can Follow Rules: Kripke's Critique of Functionalism." In Alan Berger (ed.), *Saul Kripke*. Cambridge: Cambridge University Press, pp. 343–367.

Carandini, Matteo, and David J. Heeger. 1994. "Summation and Division by Neurons in Primate Visual Cortex." *Science* 264: 1333–1336.

Chalmers, David J. 2011. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science* 12: 323–357.

Chemero, Anthony. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Chemero, Anthony, and Michael Silberstein. 2008. "After the Philosophy of Mind: Replacing Scholasticism with Science." *Journal of Philosophy* 75: 1–27.

Chirimuuta, Mazviita. 2014. "Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience." *Synthese* 191: 127–153.

Clark, Andy. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58: 7–19.

Craver, Carl F. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68: 53–74.

---. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Cummins, Robert. 1975. "Functional Analysis." *The Journal of Philosophy* 72: 741–765.

Eliasmith, Chris. 2002. "The Myth of the Turing Machine: The Failure of Functionalism and Related Theses." *Journal of Experimental and Theoretical Artificial Intelligence* 14: 1–8.

Fodor, Jerry A. 1968. "The Appeal to Tacit Knowledge in Psychological Explanations." *The Journal of Philosophy* 65: 627–640.

Godfrey-Smith, Peter. 2008. "Triviality Arguments against Functionalism." *Philosophical Studies* 145: 273–295.

Goodale, Melvyn A., and Keith G. Humphrey. 1998. "The Objects of Action and Perception." *Cognition* 67: 181–207.

Heeger, David J. 1992. "Normalization of Cell Responses in the Cat Striate Cortex." *Visual Neuroscience* 9: 181–197.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160: 106–154.

Jeannerod, Marc, and Pierre Jacob. 2005. "Visual Cognition: A New Look at the Two-Visual Systems Model." *Neuropsychologia* 43: 301–312.

Kuhl, Patricia K., and James D. Miller. 1978. "Speech Perception by the Chinchilla: Identification Functions for Synthetic Vot Stimuli." *Journal of the Acustical Society of America* 63: 905–917.

Levin, Janet. 2013. "Functionalism." In *Stanford Encyclopedia of Philosophy*, ed Edward N. Zalta. http://plato.stanford.edu/archives/fall2013/entries/functionalism/.

Levy, Arnon, and William Bechtel. 2013. "Abstraction and the Organization of Mechanisms." *Philosophy of Science* 80: 241–261.

Lewis, David K. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50: 249–258.

Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. "Perception of Speech Code." *Psychological Review* 74: 431–461.

Liberman, Alvin M., and Ignatius G. Mattingly. 1985. "The Motor Theory of Speech Perception Revised." *Cognition* 21: 1–36.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67: 1–25.

Maley, Corey, and Gualtiero Piccinini. MS. "Get the Latest Upgrade: Functionalism 6.3.1."

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Cambridge, MA: MIT Press, 2010.

McGurk, Harry, and John MacDonald. 1976. "Hearing Lips and Seeing Voices." *Nature* 264: 756–748.

Milner, A. David, and Melvyn A. Goodale. 2006. *The Visual Brain in Action. 2nd Ed.* Oxford: Oxford University Press.

Place, U.T. 1956. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47: 44–50.

Putnam, Hilary. 1967a. "The Mental Life of Some Machines." In Hector-Neri Castañeda (ed.), *Intentionality, Minds, and Perception*. Detroit: Wayne State University Press, pp. Reprinted in Putnam, 1975, *Mind, Language and Reality* (Cambridge: Cambridge University Press), pp. 408–428.

---. 1967b. "Psychological Predicates." In William H. Capitan and Daniel Davy Merrill (eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, pp. 37–48. Reprinted in Putnam, 1975, *Mind, Language and Reality* (Cambridge: Cambridge University Press), pp. 429–440. As "The Nature of Mental States.".

Rowlands, Mark. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. Cambridge, MA: MIT Press.

Rupert, Robert. 2004. "Challenges to the Hypothesis of Extended Cognition." *Journal of Philosophy* 51: 389–428.

Rust, Nicole C., and J. Anthony Movshon. 2005. "In Praise of Artifice." *Nature Neuroscience* 8: 1647–1650.

Shagrir, Oron. 2005. "The Rise and Fall of Computational Functionalism." In Yemima Ben-Menahem (ed.), *Contemporary Philosophy in Focus: Hilary Putnam*. Cambridge: Cambridge University Press, pp. 220–250.

Smart, J.J.C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68: 141–156.

Sober, Elliott. 1985. "Panglossian Functionalism and the Philosophy of Mind." *Synthese* 64: 165–193.

Sprevak, Mark. 2009. "Extended Cognition and Functionalism." *The Journal of Philosophy* 106: 503–527.

Walter, Sven. 2010. "Cognitive Extension: The Parity Argument, Functionalism, and the Mark of the Cognitive." *Synthese* 177: 285–300.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183: 313–338.

Blurring Out Cosmic Puzzles

Yann Benétreau-Dupin*

Department of Philosophy & Rotman Institute of Philosophy

Western University, Canada

Forthcoming in *Philosophy of Science*

PSA conference 2014

Abstract

The Doomsday argument and anthropic arguments are illustrations of a paradox. In both cases, a lack of knowledge apparently yields surprising conclusions. Since they are formulated within a Bayesian framework, the paradox constitutes a challenge to Bayesianism. Several attempts, some successful, have been made to avoid these conclusions, but some versions of the paradox cannot be dissolved within the framework of orthodox Bayesianism. I show that adopting an imprecise framework of probabilistic reasoning allows for a more adequate representation of ignorance in Bayesian reasoning, and explains away these puzzles.

*ybenetre@uwo.ca

1 Introduction

The Doomsday paradox and the appeal to anthropic bounds to solve the cosmological constant problem are two examples of puzzles of probabilistic confirmation. These arguments both make 'cosmic' predictions: the former gives us a probable end date for humanity, and the second a probable value of the vacuum energy density of the universe. They both seem to allow one to draw unwarranted conclusions from a *lack of knowledge*, and yet one way of formulating them makes them a straightforward application of Bayesianism. They call for a framework of inductive logic that allows one to represent ignorance better than what can be achieved by orthodox Bayesianism, so as to block these conclusions.

1.1 The Doomsday paradox

The Doomsday argument is a family of arguments about humanity's likely survival.[1] There are mainly two versions of the argument discussed in the literature, both of which appeal to a form of Copernican principle (or principle of typicality or mediocrity). A first version of the argument endorsed by, e.g., John Leslie (1990) dictates a probability shift in favor of theories that predict earlier end dates for our species, assuming that we are a typical—rather than atypical—member of that group.

The other main version of the argument, often referred to as the 'delta-t argument', was given by Richard Gott (1993) and has provoked both outrage and genuine scientific interest.[2] It claims to allow one to make a prediction about the total duration of any process of indefinite duration based only on the assumption that the moment of observation is randomly selected. A variant of this argument, which gives equivalent predictions, reasons

[1] See, e.g., (Bostrom, 2002, §6-7), (Richmond, 2006) for reviews.

[2] See, e.g., (Goodman, 1994) for opprobrium and (Wells, 2009; Griffiths and Tenenbaum, 2006) for praise.

in terms of random sampling of one's rank in a sequential process (Gott, 1994).[3] The argument goes as follows:

Let r be my birth rank (i.e., I am the r^{th} human to be born), and N the total number of humans that will ever be born.

1. Assume that there is nothing special about my rank r. Following the principle of indifference, for all r, the probability of r conditional on N is $p(r|N) = \frac{1}{N}$.

2. Assume the following improper prior probability distribution[4] for N: $p(N) = \frac{k}{N}$. k is a normalizing constant, whose value doesn't matter.

3. This choice of distributions $p(r|N)$ and $p(N)$ gives us the prior distribution $p(r)$:

$$p(r) = \int_{N=r}^{N=\infty} p(r|N)p(N)\,\mathrm{d}N = \int_{N=r}^{N=\infty} \frac{k}{N^2}\,\mathrm{d}N = \frac{k}{r}.$$

4. Then, Bayes's theorem gives us

$$p(N|r) = \frac{p(r|N)\cdot p(N)}{p(r)} = \frac{r}{N^2},$$

which favors small N.

To find an estimate with a confidence α, we solve $p(N \leq x|r) = \alpha$ for x, with $p(N \leq x|r) = \int_r^x p(N|r)\,\mathrm{d}N$. Upon learning r, we are able to make a prediction about N with a

[3] The latter version doesn't violate the reflection principle—entailed by conditionalization—according to which an agent ought to have now a certain credence in a given proposition if she is certain she will have it at a later time (Monton and Roush, 2001).

[4] As Gott (1994) recalls, this choice of prior is fairly standard (albeit contentious) in statistical analysis. It's the Jeffreys prior for the unbounded parameter N, such that $p(N)\,\mathrm{d}N \propto \mathrm{d}\ln N \propto \frac{\mathrm{d}N}{N}$. This means that the probability for N to be in any logarithmic interval is the same. This prior is called improper because it is not normalizable, and it is usually argued that it is justified when it yields a normalizable posterior.

95%-level confidence. Here, we have $p(N \leq 20r|r) = 0.95$. That is, we have:

$$p(N > 20r|r) < 5\%.$$

This result should strike us as surprising: we shouldn't be able to learn something from nothing! Indeed, according to that argument, we can make a prediction for N based only on knowing our rank r and on *not* knowing anything about the probability of r conditional on N, i.e., on being indifferent—or equally uncommitted—about any value it may take. If N is unbounded (possibly infinite), an appeal to our typical position (reflected in the choice of likelihood in the argument above) shouldn't allow us to make any prediction at all about N, and yet it does.

1.2 Anthropic reasoning in cosmology

Another probabilistic argument that claims to allow one to make a prediction from a lack of knowledge is commonly used in cosmology, in particular to solve the cosmological constant problem (i.e., explain the value of the vacuum energy density ρ_V). This parameter presents physicists with two main problems:[5]

1. The time coincidence problem: we happen to live at the brief epoch—by cosmological standards—of the universe's history when it is possible to witness the transition from the domination of matter and radiation to vacuum energy ($\rho_M \sim \rho_V$).

2. There is a large discrepancy—of 120 order of magnitudes—between the (very small) observed values of ρ_V and the (very large) values suggested by particle-physics models.

[5] See (Carroll, 2000; Solà, 2013) for an overview of the cosmological constant problem.

Anthropic selection effects (i.e., our sampling bias as observers existing at a certain time and place and in a universe that must allow the existence of life) have been used to explain both problems. In the absence of satisfying explanations, anthropic selection effects make the coincidence less unexpected, and account for the discrepancy between observations and possible expectations from available theoretical background. But there is no known reason why having $\rho_M \sim \rho_V$ should matter to the advent of life.

Weinberg and his collaborators (Weinberg, 1987, 2000; Martel et al., 1998), among others, proposed anthropic bounds on the possible values of ρ_V. Furthermore, they argued that anthropic considerations may have a stronger, predictive role. The idea is that we should conditionalize the probability of different values of ρ_V on the number of observers they allow: the most likely value of ρ_V is the one that allows for the largest number of galaxies (taken as a proxy for the number of observers).[6] The probability measure for ρ_V is then as follows:

$$\mathrm{d}p(\rho_V) = \nu(\rho_V) \cdot p_\star(\rho_V)\, \mathrm{d}\rho_V,$$

where $p_\star(\rho)\, \mathrm{d}\rho_V$ is the prior probability distribution, and $\nu(\rho_V)$ the average number of galaxies which form for ρ_V.

By assuming that there is no known reason why the likelihood of ρ_V should be special at the observed value, and because the allowed range of ρ_V is very far from what we would expect from available theories, Weinberg and his collaborators argued that it is reasonable to assume that the prior probability distribution is constant within the anthropically allowed range, so that $\mathrm{d}p(\rho_V)$ can be calculated as proportional to $\nu(\rho_V)\, \mathrm{d}\rho_V$ (Weinberg, 2000, 2). Weinberg then predicted that the value of ρ_V would be close to the mean value in that range (assumed to yield the largest number of observers). This "principle of mediocrity", as Vilenkin (1995) called it, assumes that we are typical observers.

[6] This assumption is contentious (see, e.g., (Aguirre, 2001) for an alternative proposal).

Thus, anthropic considerations not only help establish the prior probability distribution for ρ_V by providing bounds, but they also allow one to make a prediction regarding its observed value. The initial uniform distribution is turned into a prediction—a sharply peaked distribution around a preferred value—for ρ_V. This method has yielded predictions for ρ_V only a few orders of magnitudes apart from the observed value.[7] This improvement—from 120 orders of magnitude to only a few—has been seen by their proponents as vindicating anthropically-based approaches.

1.3 The problem: *Ex nihilo nihil fit*

The two examples of this section—the Doomsday argument and anthropic reasoning—share a similar structure: 1) a uniform prior probability distribution reflects an initial state of ignorance or indifference, and 2) an appeal to typicality or mediocrity is used to make a prediction. This is puzzling: these two assumptions (of indifference and typicality) are meant to express neutrality, and yet from them alone we seem to be getting a lot of information. But assuming neutrality *alone* should not allow us to learn anything!

If anthropic considerations were only able to provide us with one bound (either lower or upper bound), then the argument used to make a prediction about the vacuum energy density ρ_V would be formally identical to Gott's 1993 'delta-t argument': without knowing anything about, say, a parameter's upper bounded, a uniform prior probability distribution over all possible ranges and the appeal to typicality of the observed value favors lower values for that parameter.

I will briefly review several approaches taken to dispute the validity of the results obtained from these arguments. We will see that, because dropping the assumption of typicality isn't enough to avoid these paradoxical conclusions, it is a more adequate rep-

[7] The median value of the distribution obtained by such anthropic prediction is about 20 times the observed value ρ_V^{obs} (Pogosian et al., 2004).

resentation of ignorance or indifference that we should pursue. I wish to show that, when dealing with events we are completely ignorant about, one can use an imprecise, Bayesian-friendly framework that better handles ignorance, and avoids the paradoxical, uncomfortable consequences of the Doomsday argument, and better models the limited role anthropic considerations can play for the cosmological constant problem.

2 Typicality, indifference, neutrality

2.1 How crucial to those arguments is the assumption of typicality?

The appeal to typicality is central to Gott's 'delta-t argument', Leslie's version of the Doomsday argument, and Weinberg's prediction. This assumption has generated much of the philosophical discussion about the Doomsday paradox in particular. Nick Bostrom (2002) offered a challenge to what he calls the Self-Sampling Assumption (SSA), according to which "one should reason as if one were a random sample from the set of all observers in one's reference class." In order to avoid the consequence of the Doomsday argument, Bostrom suggested to adopt what he calls the Self-Indicating Assumption (SIA): "Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist." (*op. cit.*) But as he noted himself (Bostrom, 2002, 122-126), this SIA is not acceptable as a general principle. Indeed, as Dieks (1992) summarized:

> Such a principle would entail, e.g., the unpalatable conclusion that armchair philosophizing would suffice for deciding between cosmological models that predict vastly different chances for the development of human civilization.The infinity of the universe would become certain *a priori*.

The biggest problem with Doomsday-type arguments resting on the SSA is that their conclusion depends on the choice of reference class. What constitutes "one's reference class" seems entirely arbitrary or ill-defined: is my reference class that of all humans, mammals, philosophers, etc.? Anthropic predictions can be the object of a similar criticism: the value of the cosmological constant most favorable to the existence of life (as we know it) may not be the same as that most favorable to the existence of intelligent observers, which might be definable indifferent ways.

Relatedly, Dieks (1992) and Radford Neal (2006) showed that a careful examination of the role of indexical information in the formulation of the Doomsday argument allows one to avoid its unpleasant conclusion. In particular, Neal (2006) argued that conditionalizing on non-indexical information (i.e., all the information at the disposal of the agent formulating the Doomsday argument, including all their memories) reproduces the effects of assuming both SSA and SIA. Indeed, conditionalizing on the probability that an observer with all their non-indexical information exists (which is higher for a later Doomsday, and highest if there is no Doomsday at all) blocks the consequence of the Doomsday argument, without invoking such *ad hoc* principles, and avoids the reference-class problem.

Although full non-indexical conditioning cancels out the effects of Leslie's Doomsday argument (and, similarly, anthropic predictions), it is not clear that it also allows one to avoid the conclusion of Gott's version of the Doomsday argument. Neal (2006, 20) dismisses Gott's argument because it rests *only* on an "unsupported" assumption of typicality. There are indeed no good reasons to endorse typicality *a priori* (see, e.g., Hartle and Srednicki, 2007). One might then hope that not assuming typicality would suffice to dissolve these cosmic puzzles. Irit Maor et al. (2008) showed for instance that without it, anthropic considerations don't allow one to really make predictions about the cosmological constant, beyond just providing unsurprising boundaries, namely, that the value of the cosmological

constant must be such that life is possible.

My approach in this paper, however, will not be to question the assumption of typicality in either of these cosmic puzzles. Indeed, in Gott's version of the Doomsday paradox, we would obtain a prediction *even if we didn't assume typicality*. Consider the formulation of Gott's argument using an improper prior (§1.1 *infra*). Now, instead of assuming, a flat probability distribution for our rank r conditional on the total number of humans N $\left(p(r|N) = \frac{1}{N}\right)$, let's assume a non-uniform distribution. For instance, let's assume a distribution that favors our being born in humanity's timeline's first decile (i.e., one that peaks around $r = 0.1 \times N$). We would then obtain a different prediction for N than if we had assumed one that peaks around $r = 0.9 \times N$. This reasoning, however, yields an unsatisfying result if taken to the limit: if we assume a likelihood probability distribution for r conditional on N sharply peaked at $r = 0$, we would *still* obtain a prediction for N upon learning r, (see Fig. 1).[8]

Therefore, in Gott's Doomsday argument, we would obtain a prediction at any confidence-level, whatever assumption we make as to our typicality or atypicality, and we would even obtain one if we assume $N \to \infty$. Thus, assuming typicality or not will not allow us to avoid the conclusion of Gott's Doomsday argument. Consequently, it is toward the question of a probabilistic representation of ignorance that I will now turn my attention.

2.2 A neutral principle of indifference?

One could hope that a more adequate prior probability distribution—one that better reflects our ignorance and is normalizable—may prevent the conclusion of these cosmic puzzles (especially Gott's Doomsday argument). The idea that a uniform probability distribution is not a satisfying representation of ignorance is nothing new; this discussion is

[8]Tegmark and Bostrom (2005) used a similar reasoning to derive an upper bound on the likelihood of a Doomsday catastrophe.

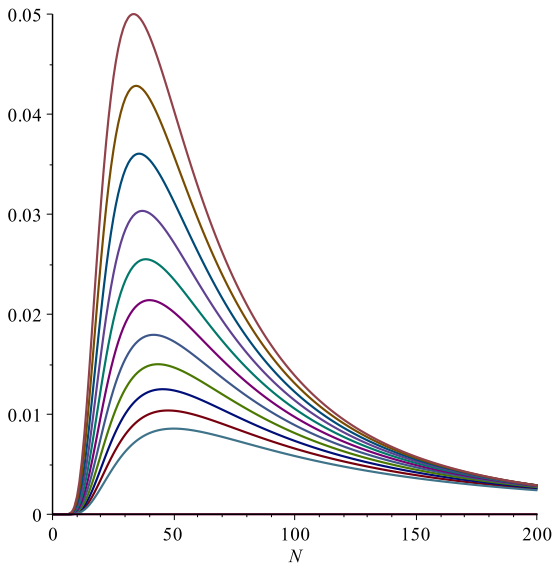


Figure 1: Posterior probability distributions for N conditional on r, obtained for $r = 100$ and assuming different likelihood distributions for r conditional on N (i.e., with different assumptions as to our relative place in humanity's timeline), which each peaks at different values $\tau = \frac{r}{N}$. The lowermost curve corresponds to a likelihood distribution that peaks at $\tau \to 0$, i.e., if we assume $N \to \infty$.

as old as the principle of indifference itself.[9] Indeed, a uniform probability distribution is unable to fulfill invariance requirements that one should expect of a representation of ignorance or indifference. As argued by John Norton (2010), a representation of ignorance or indifference

- cannot be additive (and therefore does not obey the laws of probability),
- cannot be represented by the degrees of a one-dimensional continuum, such as the reals in $[0, 1]$,
- must be invariant under redescription,

[9]See, e.g., (Syversveen, 1998) for a short review on the problem of representing non-informative priors.

- must be invariant under negation: if we are ignorant or indifferent as to whether or not α, we must be equally ignorant as to whether or not $\neg\alpha$.[10]

For instance, in the case of the cosmological constant problem, if we adopt a uniform probability distribution for the value of the vacuum energy density ρ_V over an anthropically allowed range of length μ, then we are committed to assert, e.g., that ρ_V is 3 times more likely to be found in a any range of length $\frac{\mu}{3}$ than in any other range of length $\frac{\mu}{9}$. But such an assertion is not compatible with complete ignorance as to what value ρ_V is more likely to have, hence the requirement of non-additivity for a representation of ignorance.

These criteria for a representation of ignorance or indifference cast doubt on the possibility for a probabilistic logic of induction to overcome these limitations.[11] I will argue that an imprecise model of Bayesianism, in which our credences can be fuzzy, will be able to explain away these problems, without abandoning Bayesianism altogether.

3 Dissolving the puzzles with imprecise credence

3.1 Imprecise credence

It has been argued (see, e.g., Levi, 1974; Walley, 1991; Joyce, 2010) that Bayesian credences need not have sharp values, and that there can be imprecise credences (or 'imprecise probabilities' by misuse of language). An imprecise credence model recognizes "that our beliefs should not be any more definitive or unambiguous than the evidence we have for them." (Joyce, 2010, 320)

Joyce defended an imprecise model of Bayesianism in which credences are not represented merely by a range of values, but rather by a *family* of (probabilistic) credence

[10] For an extended discussion about criteria for a representation of ignorance—with imprecise probabilities in particular—see (de Cooman and Miranda, 2007, §4-5).

[11] The same goes for improper priors, as was argued, e.g., by Dawid et al. (1973).

functions. In this imprecise probability model,

1. a believer's overall credal state can be represented by a family C of credence functions $[c_i]$ $(\dots)$. Facts about the person's opinions correspond to properties *common to all the credence functions* in her credal state.
2. If the believer is rational, then every credence function in C is a probability.
3. If a person in credal state C learns that some event D obtains $(\dots)$, then her post-learning state will be $C_D = \{c(.|D) = c(X)\frac{c(D|X)}{c(D)}, c \in C\}$.
4. A rational decision-maker with credal state C is obliged to prefer one action A to another A^* when A's expected utility exceeds that of A^* relative to *every* credence function in C. (Joyce, 2010, 288, my emphasis)

An analogy is sometimes given to illustrate this model: the overall credal state C acts as a committee whose members (each being analogous to a credence function c_i) are rational agents who do not all agree with each other and who all update their credence in the same way, by conditionalizing on evidence they all agree upon. In this analogy, the properties of the jury's opinion (the overall credal state C) are those common to *all* the committee members' opinions.

This model allows one to simultaneously represent sharp and imprecise credences, but also comparative probabilities. It can accommodate sharp credences, and then the usual condition of additivity. But it can also accommodate less sharply defined relationships when credences are fuzzy. It does so by means of a family of credence functions, each of which is treated as in orthodox Bayesianism.

This model is interesting when it comes to representing ignorance or indifference: it allows us to represent the credal state of ignorance by a *set of functions that disagree with each other.* In order to reframe our cosmic puzzles, two cases must be distinguished:

- in an unbounded case (i.e., here, Gott's Doomsday argument),[12] an imprecise prior credal set with an infinite number of probability distributions, each normalizable, will not allow one to obtain any prediction,

- in a bounded case (i.e., here, anthropic predictions for ρ_V), it is possible to construct an imprecise prior credal set with probability distributions that each favors a different value for ρ_V such that the invariance criteria given above in §2.2 are fulfilled.

3.2 Blurring out Gott's Doomsday argument: Apocalypse Not Now

Let us see how we can reframe Gott's Doomsday argument with an imprecise prior credence for the total number of humans N, or more generally for the length of any process of indefinite duration X. Let our prior credence in X, $C(X)$, be represented by a family of credal functions $\{c_\gamma\}$, each normalizable and defined on $\mathbb{R}^{>0}$. Thus, we avoid improper prior distributions. If all we assume is that X is finite but can be indefinitely large, then all we can say is that $C(X)$ is monotonically decreasing and that $\lim_{X\to\infty}(C(X)) = 0$. Let us then represent our prior credence $C(X)$ consist in the following set of functions $\{c_\gamma\}$, all of which decrease but not at the same rate (i.e., similar to a family of Pareto distributions):

$$c_\gamma(X) = \frac{k_\gamma}{X^\gamma},$$

with $\gamma > 1$ and k_γ a normalizing constant: $k_\gamma = \frac{1}{\int_0^\infty \frac{\mathrm{d}X}{X^\gamma}}$. The limiting case $\gamma \to 1$ corresponds to $X \to \infty$, but $\gamma = 1$ must be excluded to avoid a non-normalizable distribution.

If we don't want to assume anything about $\frac{\mathrm{d}C(X)}{\mathrm{d}X}$ (other than it being negative), this prior set must be such that it contains functions of decreasing rates that are arbitrarily small. That is, $\forall X \in \mathbb{R}^{>0}, \forall \epsilon \in \mathbb{R}^{<0}$, $\exists c_\gamma \in C$ s.t. $\frac{\mathrm{d}c_\gamma(X)}{\mathrm{d}X} > \epsilon$. This requirement applies

[12] The results from (Neal, 2006) to counter Leslie's Doomsday argument still apply in the imprecise framework.

not to any of the functions in C but *to the set as a whole*. It is what will block the conclusion of the Doomsday argument.[13]

Let us see how such a prior credal set avoids the conclusion of Gott's Doomsday paradox. As in the Bayesian version of the argument given in §1.1, the principle of indifference gives us an expression for the likelihood of our rank r conditional on the total number of humans N, and with our choice of prior for N, we obtain expressions for the prior for r and the posterior for N conditional on r.

The credal functions $c_\gamma(r)$ in the set of distributions for the prior credence in r, $C(r)$ can be expressed as follows:

$$c_\gamma(r) = \int_{N=r}^{N=\infty} p(r|N) \cdot c_\gamma(N)\,\mathrm{d}N = \int_{N=r}^{N=\infty} \frac{k_\gamma}{N^{\gamma+1}}\,\mathrm{d}N.$$

Bayes' theorem then yields an expression for posterior credal functions:

$$c_\gamma(N|r) = \frac{p(r|N) \cdot c_\gamma(N)}{c_\gamma(r)} = \frac{k_\gamma}{N^{\gamma+1} \cdot \int_{N=r}^{N=\infty} \frac{k_\gamma}{N^{\gamma+1}}\,\mathrm{d}N}.$$

To find a prediction for N with a 95%-level confidence, we solve $C(N \leq x|r) = 0.95$ for x, with $C(N \leq x|r) = \int_r^x C(N|r)\,\mathrm{d}N$. Now, as $\gamma \to 1$, the prediction for x such that $C(N \leq x|r) = 95\%$ diverges. In other words, this imprecise representation of prior credence in N, reflecting our ignorance about N, does not yield any prediction about N (see figure 2).

Any of the credal functions c_γ in the credal set as defined here would yield a prediction if taken individually. However, it is clear that this prediction would rest solely on an arbitrary choice of prior that doesn't reflect our initial state of ignorance. Without the possibility for my prior credence to be represented not by a single probability distribution

[13] In order to avoid too sharply peaked distributions (at $X \to 0$), further constraints can be placed on the variance of the distributions (namely, an lower bound on the variance), without it affecting my argument.

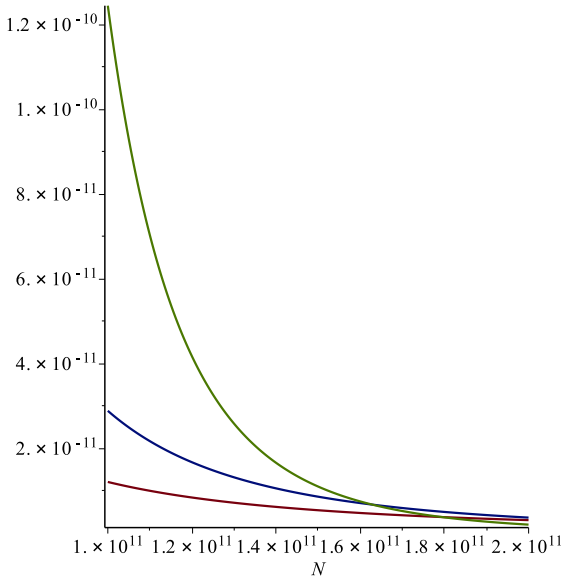


Figure 2: Posterior probability distributions for N conditional on r, obtained for $r = 1.2 \cdot 10^{11}$ and assuming different prior distributions for N (i.e., with different assumptions as to the total number of humans there will ever be).

but by an *infinite set* of probability distributions, I cannot avoid obtaining an arbitrarily precise prediction.

Other distributions that decrease at different rates (i.e., not as inverse powers of N) could have been included in the prior credal set $\{c_\gamma\}$, as long as they fulfill the criteria listed at the beginning of this section. However, no other distribution we could include would change this conclusion. In order to represent our credence about the length of a process of indefinite duration, it is necessary that our prior credal set includes the functions c_γ defined earlier, and that is sufficient to avoid the conclusions of the Doomsday argument.

3.3 Blurring out anthropic predictions

We are ignorant about what value of the vacuum energy density ρ_V we should expect from our current theories. We now want to express the fact that, in the absence of a prior credence that tells us something about what we should expect, we shouldn't be in a position to confirm or not the assumption of typicality on which anthropic predictions for the cosmological constant rest.

If we substitute imprecise prior and posterior credences in the formula from (Weinberg, 2000, see §1.2 *infra*), we have:

$$\mathrm{d}C(\rho_V) = \nu(\rho_V) \cdot C_\star(\rho_V)\, \mathrm{d}\rho_V,$$

with $C_\star(\rho_V)$ a prior credal set that will exclude all values of ρ_V outside the anthropic bounds, and $\nu(\rho_V)$ the average number of galaxies which form for ρ_V, which as in §1.2 peaks around the mean value of the anthropic range. In order for the prior credence $C_\star$ to express our ignorance, it should be such that it doesn't favor any value of ρ_V.

With the imprecise model, such a state of ignorance can be expressed by a set of probability distributions $\{c_{\star i}(\rho_V)\}$, all of which normalizable over the anthropic range and such that $\forall \rho_V, \exists c_{\star i}, c_{\star j} \in C_\star$ such that ρ_V is favored by $c_{\star i}$ and not by $c_{\star j}$.[14] Such a prior credal set will not favor any value of ρ_V. Moreover, in order to fulfill the criterion of invariance under negation (according to which $C_\star(\rho_V) = C_\star(\neg\rho_V)$, see §2.2), one could define a credal set representing ignorance to be such that $\forall \rho_V, \forall c_{\star i} \in C_\star, \exists c_{\star j} \in C_\star$ such that $c_{\star i} = 1 - c_{\star j}$.[15]

With a prior credal set $C\star$ thus defined, even with a distribution $\nu(\rho_V)$ peaked around

[14] This can be obtained, for instance, by a family of Gamma distributions, each of which giving an expected value at a different point in the anthropically allowed range. As in §3.1, In order to avoid dogmatic functions, a lower bound can be placed on the variance of all the functions in $C_\star$.

[15] But such a symmetry requirement need not be required in all cases; unwarranted conclusions can be avoided without necessarily assuming this condition.

the mean value of that anthropic range, the prediction $C(\rho_V)$ becomes very imprecise all over the anthropic range. But more importantly, we won't have $C(\rho_V^{obs}) > C_\star(\rho_V^{\text{obs}})$, i.e., there will be no agreement among *all* the distributions $c_{\star i} \in C_\star$ that learning the actual value ρ_V^{obs} will provide a confirmatory boost for our assumption of typicality.

The imprecise model can then provide us with a way to express our ignorance such that our assumption of typicality is neither confirmed nor disconfirmed. And yet, that same approach doesn't prevent Bayesian induction altogether. Indeed, all the functions in $C_\star$ being probability distributions that can be treated as in orthodox Bayesianism, any of them can be updated and, in principle, converge toward a sharper credence, provided sufficient updating.

4 Conclusion

These cosmic puzzles show that, in the absence of an adequate representation of ignorance, a logic of induction will inevitably yield unwarranted results. Our usual methods of Bayesian induction are ill-equiped to allow us to address both puzzles. I have shown that the imprecise credence framework allows us to treat both arguments in a way that avoids their undesirable conclusions. The imprecise model rests on Bayesian methods, but it is expressively richer than the usual Bayesian approach that only deals with single probability distributions (i.e., sharp credence functions).

Philosophical discussions about the value of the imprecise model usually center around the difficulty to define updating rules that don't contradict general principles of conditionalization (especially the problem of dilation). But the ability to solve such paradoxes of confirmation and avoid unwarranted conclusions should be considered as a crucial feature of the imprecise model and play in its favor.

References

Aguirre, A. (2001, September). Cold Big-Bang Cosmology as a Counterexample to Several Anthropic Arguments. *Physical Review D 64*(8), 1–12.

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

Carroll, S. (2000). The Cosmological Constant. *arXiv: astro-ph/0004075v2*, 1–50.

Dawid, A. P., M. Stone, and J. V. Zidek (1973). Marginalization Paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society. Series B (Methodological) 35*(2), 189–233.

de Cooman, G. and E. Miranda (2007). Symmetry of models versus models of symmetry. In W. L. Harper and G. Wheeler (Eds.), *Probability and Inference. Essays in Honour of Henry E. Kyburg Jr*, Number April, pp. 67–149. London: College Publications.

Dieks, D. (1992). Doomsday–Or: The Dangers of Statistics. *The Philosophical Quarterly 42*(166), 78–84.

Goodman, S. N. (1994). Future Prospects Discussed. *Nature 368*(March), 108–109.

Gott, J. R. (1993). Implications of the Copernican Principle for our Future Prospects. *Nature 363*(6427), 315–319.

Gott, J. R. (1994). Future Prospects Discussed. *Nature 368*(March), 108.

Griffiths, T. L. and J. B. Tenenbaum (2006). Optimal Predictions in Everyday Cognition. *Psychological Science 17*(9), 767–773.

Hartle, J. and M. Srednicki (2007, June). Are We Typical? *Physical Review D 75*(12), 123523.

Joyce, J. M. (2010). A Defense of Imprecise Credences in Inference and Decision Making. *Philosophical Perspectives 24*(1), 281–323.

Leslie, J. A. (1990). Is the End of the World Nigh? *The Philosophical Quarterly 40*(158), 65–72.

Levi, I. (1974). On Indeterminate Probabilities. *The Journal of Philosophy 71*(13).

Maor, I., L. Krauss, and G. Starkman (2008, January). Anthropic Arguments and the Cosmological Constant, with and without the Assumption of Typicality. *Physical Review Letters 100*(4), 041301.

Martel, H., P. R. Shapiro, and S. Weinberg (1998). Likely Values of the Cosmological Constant. *The Astrophysical Journal 492*(1), 29–40.

Monton, B. and S. Roush (2001). Gott's Doomsday Argument. *http://philsci-archive.pitt.edu/id/eprint/1205*, 1–23.

Neal, R. M. (2006). Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning. *Arxiv preprint math/0608592* (0607), 1–56.

Norton, J. D. (2010). Cosmic Confusions: Not Supporting versus Supporting Not. *Philosophy of Science 77*(4), 501–523.

Pogosian, L., A. Vilenkin, and M. Tegmark (2004, July). Anthropic Predictions for Vacuum Energy and Neutrino Masses. *Journal of Cosmology and Astroparticle Physics 7*(005), 1–17.

Richmond, A. (2006). The Doomsday Argument. *Philosophical Books 47*(2), 129–142.

Solà, J. (2013, August). Cosmological Constant and Vacuum Energy: Old and New Ideas. *Journal of Physics: Conference Series 453*(1), 012015.

Syversveen, A. R. (1998). Noninformative Bayesian Priors. Interpretation and Problems with Construction and Applications.

Tegmark, M. and N. Bostrom (2005, December). Is a Doomsday Catastrophe Likely? *Nature 438*(7069), 754.

Vilenkin, A. (1995). Predictions from Quantum Cosmology. *Physical Review Letters 74*(6), 4–7.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Weinberg, S. (1987). Anthropic Bound on the Cosmological Constant. *Physical Review Letters 59*(22), 2607–2610.

Weinberg, S. (2000). A Priori Probability Distribution of the Cosmological Constant. *arXiv preprint astro-ph/0002387*, 0–15.

Wells, W. (2009). *Apocalypse When? Calculating How Long the Human Race Will Survive*. Springer Praxis Books. Praxis.

COARSE-GRAINING AS A ROUTE TO MICROSCOPIC PHYSICS: THE RENORMALIZATION GROUP IN QUANTUM FIELD THEORY

BIHUI LI

ABSTRACT. The renormalization group (RG) has been characterized as merely a coarse-graining procedure that does not illuminate the microscopic content of quantum field theory (QFT), but merely gets us from that content, as given by axiomatic QFT, to macroscopic predictions. I argue that in the constructive field theory tradition, RG techniques do illuminate the microscopic dynamics of a QFT, which are not automatically given by axiomatic QFT. RG techniques in constructive field theory are also rigorous, so one cannot object to their foundational import on grounds of lack of rigor.

Date: June 24, 2014.

1. Introduction

The renormalization group (RG) in quantum field theory (QFT) has received some attention from philosophers for how it relates physics at different scales and how it makes sense of perturbative renormalization (Huggett and Weingard 1995; Bain 2013). However, it has been relatively neglected by philosophers working in the axiomatic QFT tradition, who take axiomatic QFT to be the best vehicle for *interpreting* QFT. Doreen Fraser (2011) has argued that the RG is merely a way of getting from the microscopic principles of QFT, as provided by axiomatic QFT, to macroscopic experimental predictions. Thus, she argues, RG techniques do not illuminate the theoretical content of QFT, and we should stick to interpreting axiomatic QFT. David Wallace (2011), in contrast, has argued that the RG supports an effective field theory (EFT) interpretation of QFT, in which QFT does not apply to arbitrarily small length scales. Like Wallace, physicists generally regard the RG to be foundationally significant, as recent QFT textbooks indicate (Zee 2010; Duncan 2012).

My main objective is to question Fraser's claims that the RG is *only* a way to get from the microscopic principles of QFT to macroscopic predictions, and that it has no significance for the theoretical content of QFT. Unlike Wallace, I do this without endorsing an EFT interpretation of QFT. Instead, I elucidate the foundational significance of the RG by describing its role in determining whether various Lagrangians could possibly describe QFTs living on continuous spacetime—that is, whether these Lagrangians are well-defined in the ultraviolet (UV) limit. This problem is an important one in the foundations of QFT and it is the central aim of constructive field theory, which attempts to construct interacting models of QFT satisfying certain axioms. The existence of the UV limit is relevant to whether we should interpret a particular Lagrangian as describing an EFT or as potentially applicable to all length scales, so if the RG helps determine the existence of this limit, then the RG is significant for the interpretation of QFT.

To forestall the objection that RG methods are not rigorous enough for philosophical attention, I look at the RG as used in constructive field theory, a tradition that philosophers take to be rigorous. Many in this tradition use RG methods to determine whether various Lagrangians have a well-defined UV limit. The rigor of these RG methods as compared to the RG methods that physicists typically use lies in the employment of well-controlled approximations rather than ill-controlled approximations.

My plan is as follows. In the next section, I provide more specifics on the various theoretical approaches to QFT and flesh out the claims that I have attributed to Fraser. In Section 3, I sketch the formalism of perturbative QFT, describing the problems that constructive QFT aims to solve. In Section 4, I sketch the "physicists' version" of the RG as a pedagogical attempt to show how the RG can answer the question of whether a UV limit for a given Lagrangian exists. In Section 5, I explain how constructive field theory tries to resolve the problems with perturbative QFT and how it attempts to fill in the mathematical gaps in the physicists' version of the RG. In doing so, I sketch how constructive field theory uses RG methods to try to construct models of QFT that exist in continuous spacetime. I conclude by musing on what the argument of this paper implies about the relationship between the various theoretical strands of QFT.

2. The Debate So Far

In the early days of QFT, physicists ran into a host of mathematical pathologies such as divergences in their perturbation expansions. To get around these, they deployed calculational methods such as perturbative renormalization without fully understanding why these methods worked.

Axiomatic QFT grew out of attempts to make the mathematical character of QFT clearer. One variant of axiomatic QFT is algebraic QFT, which I will not discuss here.

Instead, I focus on the Wightman axioms or the Osterwalder-Schrader (OS) axioms, which specify the properties that a theory's Wightman functions or Schwinger functions, respectively, must satisfy to define a QFT.[1] However, these properties are insufficient to define a QFT's dynamics. For more dynamical details, we turn to constructive QFT (CQFT),[2] which attempts to construct specific interacting models of QFT that satisfy the OS axioms. Such models, if they exist, automatically satisfy the Wightman axioms, according to the Osterwalder-Schrader reconstruction theorem (Rivasseau 1991). CQFT takes its models of interest to be those characterized by Lagrangians that physicists use. One of the aims of CQFT is to find out if these Lagrangians correspond to non-trivial QFTs in the UV limit.

A QFT that satisfies either set of axioms must have a UV limit: effective field theories violate the axiom of postivity in the OS axioms. A typical approach in CQFT is to start with a lattice QFT or an effective field theory with a momentum cutoff, and then to figure out what happens to the Lagrangian at a fixed momentum scale when the lattice spacing is taken to zero, or when the cutoff is taken to infinity. If the model that results when this limit is taken is trivial (all the coupling constants in the Lagrangian go to zero) or ill-defined (some coupling constant becomes infinite in the limit), then one concludes that there does not exist a non-trivial model of that QFT in continuum spacetime.

The RG was first developed in an unrigorous manner within perturbative QFT. It provides an account of how the dynamics of QFTs change with length or energy scale. These changes are manifested as changes in the value of the coupling parameters in a theory's Lagrangian. Part of the importance of the RG lies in how it explains the empirical success of perturbative renormalization. The RG provides a physical picture of why one has to change the values of coupling parameters in order to avoid divergences. As mentioned earlier, physicists have generally regarded the RG to be foundationally and interpretively significant.

[1] The Schwinger and Wightman functions are important because any observable can be computed from them.
[2] Also known as *constructive field theory*.

In contrast, there is a refrain among philosophers along the lines sketched by Fraser (2011, 131):

> RG methods make a significant contribution to the articulation of the empirical content of QFT and to clarifying the nature of the relationship between the empirical and the theoretical content. However, RG methods do not shed light on the theoretical content of QFT. For this reason, appeal to RG methods does not decide the question of which set of theoretical principles are appropriate for QFT... The reason that constructive field theorists are able to exploit RG methods—even though they reject elements of the theoretical content of LQFT—is that RG methods concern the empirical structure of the theory rather than the theoretical content.

In a similar vein, Kuhlmann, Lyre, and Wayne (2002) characterize the RG as providing "a deductive link between fundamental QFT and experimental predictions". This echoes the thought, latent in Fraser's writings, that there is some "fundamental QFT" given prior to using the RG, presumably by some axiomatic form of QFT, and that all the RG does is link this fundamental theory to experimental predictions. Fraser takes this thought to undercut Wallace's argument that RG methods support a particular interpretation of QFT.

This pattern of reasoning is common in the philosophy of physics: for foundational or interpretive purposes, we should focus on only the "fundamental principles" of a theory, given by its axioms, because these constitute the entire theoretical content of the theory. Methods to extract predictions from these principles add no new theoretical content, only pragmatic filigree.

However, as I shall argue, RG methods do have foundational significance because they are one of the main ways in which CQFT proves the existence or non-existence of models of QFT satisfying the OS axioms. Thus, they bear on the interpretively relevant question of

whether certain models of QFT can exist in continuous spacetime. Furthermore, there exist rigorous ways to implement the RG, and these are used in CQFT.

3. Perturbative Quantum Field Theory

CQFT arose out of a need to mathematically justify perturbative QFT. In much of QFT, perturbative renormalization is a key technique for deriving finite results for empirically measurable quantities like scattering cross-sections. A first pass at calculating these quantities leads to divergent terms in the relevant perturbation expansions. Perturbative renormalization adjusts the coupling parameters so as to remove these divergent terms. However, this procedure, as presented in introductory QFT textbooks, is carried out on a purely formal basis. While the procedure is justified in one sense by its empirical success, they are not justified by a mathematical understanding of the nature of the perturbative expansion. One of the aims of CQFT is to justify these rules mathematically. In the rest of this section, I offer a brief sketch of perturbative renormalization in a simple case so as to illustrate the room for justification that CQFT tries to provide.

One quantity of central importance in QFT is the partition function, which is defined in terms of the Lagrangian $\mathcal{L}(\phi)$ as follows:

$$Z = \int \mathcal{D}\phi e^{\int \mathcal{L}(\phi) d^4x} \tag{1}$$

Here I have assumed four dimensions for the purpose of the example. The "$\mathcal{D}$" indicates that this integral is a functional integral, sometimes called a Feynman path integral. Intuitively, the integration ranges over the space of "possible functions" ϕ, for some value of "possible".[3] Path integrals also feature in expressions for the Green's functions, which are closely related to experimental measurements.

[3] As we will see later, one of the first tasks of constructive field theory is to give a precise meaning to the measure $\mathcal{D}\phi$.

These path integrals can be given a straightforward finite, analytic expression when the action involved is that of a free scalar field with no interactions. In this case, $\mathcal{L} = \frac{1}{2}\left(\left(\partial\phi^2\right)^2 - m^2\phi^2\right)$. For interacting fields, physicists typically use perturbation theory to evaluate the path integrals. Since the path integral for the free field has a known analytic expression, the perturbations are applied using the free field case as a reference—we consider the interaction as a small perturbation to the free field Lagrangian. The following example illustrates how this is done in a simple case.

Suppose a small interaction $-\frac{\lambda}{4!}\phi^4$ is added to the free field Lagrangian, so that $\mathcal{L} = \frac{1}{2}\left(\left(\partial\phi^2\right)^2 - m^2\phi^2\right) - \frac{\lambda}{4!}\phi^4$. This is the Lagrangian of the so-called ϕ^4 theory, which describes a self-interacting scalar field. The partition function is

$$Z = \int \mathcal{D}\phi e^{\int d^4x\left(\left(\left(\partial\phi^2\right)^2 - m^2\phi^2\right) - \frac{\lambda}{4!}\phi^4\right)}.$$

Assuming λ to be small, we then convert the $e^{-\frac{\lambda}{4!}\phi^4}$ factor into a Taylor series in λ:

(2)

$$Z = \int \mathcal{D}\phi \left(1 - \frac{\lambda}{4!}\int_{x_1} \phi^2(x_1)\,dx_1 + \frac{1}{2}\left(\frac{\lambda}{4!}\right)^2 \int_{x_1,x_2} \phi^4(x_1)\,\phi^4(x_2)\,dx_1 dx_2 + \cdots\right) e^{\int d^4x\left(\left(\partial\phi^2\right)^2 - m^2\phi^2\right)}$$

where I have included only the first two terms of the Taylor series to illustrate the general rule.

Unlike in the free field case, when evaluating path integrals such as the above, some of the individual terms in the Taylor series are infinite. These divergences make it difficult to directly compute experimentally measurable quantities such as scattering cross-sections from the path integral. In many cases, the divergences can be removed by the process of perturbative renormalization. This process starts with regularization, a way of eliminating the influence of high-momenta processes which cause the divergences, and, for some methods of regularization, the addition of counterterms to compensate for regularization. Regularization is typically followed by renormalization, which consists of rewriting the Lagrangian and

expressions for quantities like cross-sections in terms of "renormalized" coupling parameters rather than the "bare" parameters that we started with. These methods have proven to be empirically successful for theories like quantum electrodynamics.

Even though perturbative renormalization removes the term-by-term divergences that occur in (2), they leave unresolved other issues. It is suspected that expansions like (2) do not converge and are at best asymptotic. An asymptotic series can be useful if we know which function the series is asymptotic to, but perturbative QFT on its own does not provide this information. Part of the CQFT program involves showing that some properties of the non-perturbative solutions to the equations of motion guarantee that certain methods of summing asymptotic perturbative expansions will lead to a unique solution. I will not discuss this part of the CQFT program. The part I will discuss in Section 5.2 involves using the RG to evaluate (1). Here the problem of divergent perturbation series manifests itself as the so-called large field problem, which will also be addressed in Section 5.2.

The other problem with perturbative QFT that CQFT tries to resolve is a proper definition of the measure of (1). Again, we will see in Section 5.2 how this is done in CQFT. For now, I move on to discussing how the RG is important not just as a way to calculate empirical quantities, but also to determine whether a given Lagrangian exists in the UV limit.

4. The Renormalization Group

The RG explains perturbative renormalization non-perturbatively. It gives an account of changing coupling parameters that is not based wholly on formal perturbative series and perturbative renormalization. The RG is widely used in the non-rigorous variants of QFT used by physicists and in constructive field theory. For convenience, I follow Wallace (2011) in calling the former "conventional QFT". While the constructive field theory treatment of the RG plugs many mathematical gaps in conventional QFT, the important conceptual insights are

already present in the conventional treatment. The conventional understanding of fixed points and RG flows suffices to help us understand how RG techniques give us not just macroscopic information, but also information about the existence of a UV limit. Here, I sketch the RG as typically presented conventionally, explain its significance for foundational questions, and point to the places where a constructive treatment might fill in some gaps. I leave the constructive treatment to Section 5.2.

The RG is a particularly effective way of computing the partition function (1). Intuitively, the operation of an RG transformation is often described as integrating out high-momentum degrees of freedom to obtain an effective action over the remaining low momenta. Formally, this transformation is often written as follows:

$$\int \mathcal{D}\phi_L \int \mathcal{D}\phi_H e^{iS[\phi_H,\phi_L]} = \int \mathcal{D}\phi_L e^{iS_\Lambda[\phi_L]}, \tag{3}$$

where ϕ_L indicate field configurations whose Fourier transforms have support over momenta less than Λ, and ϕ_H indicate field configuations whose Fourier transforms have support over momenta more than Λ. $S_\Lambda[\phi_L]$ is known as the effective action because it "acts like the full action" $S[\phi_H,\phi_L]$ but involves fewer degrees of freedom. It behaves like the full action when we describe our system with a reduced set of variables, that is, with only ϕ_L instead of $\phi_L+\phi_H$. This strategy of using effective actions at lower momentum scales to help evaluate the full integral is important in constructive field theory and in less rigorous work within QFT. Roughly speaking, RG methods proceed by doing many such integrations over infinitesimal momentum shells. This is a more effective way of computing the partition function compared to methods that try to integrate over all momenta at once, because many of the expansions that we have found to be helpful in evaluating the partition function are effective only at fixed momentum scales.

Denote the transformation (3), taking a more fine-grained action to a more coarse-grained action, by $\mathcal{R}$. The more times we iterate $\mathcal{R}$ on the action S, the larger the range of momenta we can integrate over. Each application of $\mathcal{R}$ changes the coupling parameters of terms in the action.[4] That is, each $\mathcal{R}$ moves S along a trajectory in the space of actions. Sometimes this flow can end up in a fixed point: a point where the transformation maps the action defined by that point in the space of actions to itself. That is, a fixed point is a point S where $\mathcal{R}S = S$.

The existence of a fixed point is important for determining if a given Lagrangian has a UV limit. A continuum theory exists if at an arbitrary fixed momentum scale Λ_L, the effective action S_{Λ_L} that we calculate using RG transformations converges as the momentum cutoff goes to infinity. That is, suppose we have calculated S_{Λ_L} by iterating $\mathcal{R}$ many times on an initial action $S_{\Lambda_{UV}}$, where Λ_{UV} is a momentum scale higher than Λ_L. We then see what happens to the S_{Λ_L} that we calculate with iterated $\mathcal{R}$s as we *increase* Λ_{UV}. The theory associated with S_{Λ_L} has a UV limit if there is some S for which $\lim_{\Lambda_{UV}\to\infty} S_{\Lambda_L} = S$. That is, it has a UV limit if, as we raise Λ_{UV} and have to repeat $\mathcal{R}$ more and more times in order to compute S_{Λ_L} from increasingly fine-grained actions, we get a stable result for S_{Λ_L}, showing the existence of a fixed point. In this way, the existence of fixed points of a certain sort can help us answer the question about whether various models of QFT can exist in contiuous spacetime.

Importantly, even though it is true that $\mathcal{R}$ only takes us from a more fine-grained, microscopic action to a more coarse-grained, macroscopic action, it is nevertheless the case that we can use $\mathcal{R}$ to determine whether a UV limit exists, by way of the fixed point analysis just described. This reveals the mistake in Fraser's claim that $\mathcal{R}$, as a coarse-graining

[4]This includes the possibility that terms that didn't exist before gain a non-zero coefficient under the transformation.

procedure, can only be a tool to get from the microscopic principles to macroscopic predictions and not a way to illuminate the microscopic content of the theory.

Indeed, in general, the methods used in constructive field theory to determine whether a given Lagrangian exists in the continuum limit all rely on some kind of multiscale analysis for problems with spacetime dimension $D \geq 3$ (Douglas 2011). The phase space analysis of Glimm and Jaffe (1987) is another example of such a multiscale analysis. The importance of the RG and phase space analysis in finding continuum solutions of QFT shows that the fact that a mathematical method implements some kind of scaling does not imply that it is merely a way to get from an already given microscopic physics to a merely "phenomenological" macroscopic physics.

5. Constructive Field Theory and the Renormalization Group

While we saw in the previous section how the RG as expressed in conventional QFT sheds light on the existence of UV limits, constructive field theory distinguishes itself from other means of finding a UV limit by its greater rigor. This rigor consists in:

(1) Making sure that the relevant functional integrals are well-defined;

(2) In computing the functional integrals, making sure that the approximations and expansions used are well-controlled.

I illustrate point 1 in Section 5.1 and point 2 in Section 5.2.

5.1. **Functional Integrals in Constructive Field Theory.**

I now sketch the constructive field theory approach to defining functional integrals. For simplicity, I consider the ϕ^4 theory (with dimension unspecified for now). Constructive field theorists like to operate with Euclidean functional integrals because this allows them to use the theory of Gaussian integrals. Much of the work in defining (1) draws from this probability theory basis. In Euclidean field theory, we can regard the real-valued fields $\phi(x)$ as random variables on the d-dimensional Euclidean

space $\mathbb{R}^d$. These random variables are associated with a Gaussian measure that is perturbed by an interaction term. The Gaussian measure is associated with the properties of free particles, and the interaction term with interactions between particles.

The Gaussian random field $\phi(x)$ has a mean given by $\int \phi(x) d\mu_C(\phi) = 0$ and a covariance given by $\int \phi(x)\phi(y) d\mu_C(\phi) = (-\Delta + m^2)^{-1}(x,y) \equiv C(x,y)$. We can formally write $C(x,y) = \int_{\mathbb{R}^d} \frac{e^{ip(x-y)}}{p^2+m^2} dp$, which will help us understand ultraviolet regularization later. The Schwinger functions $\langle F(\phi) \rangle$ can be formally written as

$$\langle F(\phi) \rangle = \frac{1}{Z} \int F(\phi) e^{-V(\phi)} d\mu_C(\phi), \tag{4}$$

where $Z = \int e^{-V(\phi)} d\mu_C(\phi)$. In the case of ϕ^4 theory, $V(\phi) = \lambda \int_{\mathbb{R}^d} \phi(x)^4 dx$, where λ is a coupling parameter.

The first task of constructive field theory is to modify the above expression for $\langle F(\phi) \rangle$ so that it is well-defined. The measure $d\mu_C(\phi)$ is generally not well-defined before the following steps: ultraviolet regularization, infrared regularization, and, in four dimensions, the addition of counterterms.[5] Ultraviolet regularization is required to ensure that the product of distributions $\phi(x)^4$ is well-defined. This is usually done through a momentum cutoff or lattice regularization. For brevity's sake, I outline only the momentum cutoff method. The momentum cutoff is imposed by altering $C(x,y)$ to $C_\varepsilon(x,y) = \int_{\mathbb{R}^d} \frac{e^{ip(x-y)}}{p^2+m^2} e^{-\varepsilon|p|^2} dp$, $\varepsilon > 0$. Infrared regularization imposes a finite volume Λ over which the integral for $V(\phi)$ is to be carried out. So $V(\phi)$ becomes $V_\Lambda(\phi) = \lambda \int_\Lambda \phi(x)^4 dx$. Finally, if $d = 4$, we have to add a counterterm $\delta V_{\Lambda,\varepsilon}$ to V_Λ, so we have $V_{\Lambda,\varepsilon} = V_\Lambda + \delta V_{\Lambda,\varepsilon}$ in the exponent instead.[6]

[5] In two or three dimensions, the ϕ^4 model is superrenormalizable and no counterterms are needed.

[6] I leave out the details of the form of $\delta V_{\Lambda,\varepsilon}$ for brevity. See Watanabe (2000) for details.

The upshot of all this is that the formal expression (4) is turned into a well-defined expression:

$$\langle F(\phi)\rangle_{\Lambda,\varepsilon} = \frac{1}{Z_{\Lambda,\varepsilon}} \int F(\phi) e^{-V_{\Lambda,\varepsilon}(\phi)} d\mu_{C_\varepsilon}(\phi). \tag{5}$$

The task of constructive field theory is to show that this expression has a well-defined limit as $\varepsilon \to 0$ and $\Lambda \to \infty$. If this limit exists, then the Lagrangian in question has a UV limit. Multiscale methods allow one to evaluate the integral by decomposing it into momentum scale-indexed parts. This decomposition allows for each scale-indexed part to be evaluated using certain kinds of expansions, without running into problems with the expansions failing when they try to cover too large a momentum range. The RG is one such multiscale method, and we will now see how it works in constructive field theory.

5.2. **Applying the Renormalization Group Rigorously.** In Section 4 we saw a sketch of the physical ideas behind the RG. Constructive field theorists implement the same ideas using more rigorous mathematics. As with more cavalier implementations of the RG, the existence of a UV limit in constructive field theory is linked to the existence of fixed points of RG transformations. However, many RG methods used in conventional QFT fail to account for the large field problem. Many non-perturbative approaches to the RG make use of non-perturbative approximations that we do not know how to place error bounds on.[7]

Constructive field theory tries to find the UV limit using approximations that are better controlled than those of conventional QFT. One way to do this is via the exact renormalization group (ERG).[8] The term "exact" in this context indicates that the RG is implemented

[7]For example, this a defect of the "functional renormalization group" tradition, as Gurau, Rivasseau, and Sfondrini (2014) point out.

[8]Note of caution: some who work in the tradition of the functional renormalization group take themselves to be using the "exact" renormalization group, which they take to a term referring to Wilson's non-perturbative understanding of RG flows (Rosten 2012; Bagnuls and Bervillier 2001). However, as explained previously, the lack of precise error bounds on their approximations sets them apart from the constructive field theory tradition.

non-perturbatively and that the approximations involved are well-controlled. Benfatto, Cassandro, Gallavotti, Nicoló, Olivieri, Presutti, and Scacciatelli (1980), Gawędzki and Kupiainen (1983), Gawędzki and Kupiainen (1985), Brydges, Dimock, and Hurd (1995), and Abdesselam (2007) are examples of how the ERG is used in constructive field theory. I now sketch an RG analysis based on integrating out fluctuations over slices of momentum space, showing how one may determine whether a given Lagrangian has a UV limit in this way.[9]

As mentioned in Section 4, the basic idea of the RG is to integrate the functional integral over momentum slices. This avoids the failures of various kinds of expansions when one integrates over a large range of momenta in one step. In the constructive field theory framework this integration can take place by dividing the covariance C_ε into parts that correspond to momentum slices. Notating C_ε as D for convenience, we have

$$D = \sum_{k=0}^{N} D_k,$$

with independent Gaussian variables $\phi_k(x)$ that each have mean 0 and covariance D_k. Each ϕ_k corresponds to a fluctuation field of momentum scale L^k. The slices of measure D_k are defined as follows:

$$D_k(x,y) = \int_{\mathbb{R}^d} \frac{e^{ip(x-y)}}{p^2+m^2}(\chi(L^{-k}) - \chi(L^{-(k-1)}p))dp, \quad k = 1,2,\ldots,N,$$

$$D_0(x,y) = \int_{\mathbb{R}^d} \frac{e^{ip(x-y)}}{p^2+m^2}\chi(p)dp,$$

where $\chi(p) = e^{-p^2}$ serves as a cutoff function. The D_k serve the purpose of scale decomposition because each D_k effectively isolates the range of momenta between L^{k-1} and L^k.

[9]Besides momentum slice integration, another way of implementing the RG in constructive field theory is the block spin transformation, where one treats the quantum field in a lattice setting.

Defining $H(\phi) \equiv H_N(\phi) = e^{-V_{\Lambda,\varepsilon}(\phi)}$, $\phi_{k,0} = \sum_{j=0}^{k} \phi_j$, and $D_{k,0} = \sum_{j=0}^{k} D_j$, $k = 0, 1, \ldots, N$, we can define the operation of scaling out higher momenta as follows:

$$H_{k-1}(\phi_{k-1,0}) = \int d\mu_{D_k}(\phi_k) H_k(\phi_k + \phi_{k-1,0}), \quad k = N, N-1, \ldots, 1. \tag{6}$$

H_{k-1} is simply the coarse-grained version of H_k, with the higher momenta integrated out. In an RG analysis, we would want to iterate this operation of integrating out higher momenta. Before iterating it, however, we rescale the field $\phi_{k-1,0}$ so that it has a wavelength comparable to ϕ_k's. The rescaled field is defined as $\tilde{\phi}_k(x) = L^{-k(d-2)/2}\phi_k(L^{-k}x)$. We also rescale the covariance D_k, the details of which I omit for brevity.[10] Then we define the rescaled H_k by

$$\tilde{H}_k(\tilde{\phi}_{k,0}) = H_k(\phi_{k,o}).$$

This gives us the RG transformation

$$\tilde{H}_{k-1}(\tilde{\phi}_{k-1,0}) = \int d\mu_{\tilde{D}_k}(\tilde{\phi}_k) \tilde{H}_k(\tilde{\phi}_k(\cdot) + L^{-(d-2)/2}\tilde{\phi}_{k-1,0}(L^{-1}\cdot)).$$

While we have been using the notation $H(\phi) = e^{-V_{\Lambda,\varepsilon}(\phi)}$ for convenience, we can think of the RG transformation as acting on the action V. Each transformation consists of the following steps:

(1) Rescaling of the fields;

(2) Integrating over a momentum slice;

(3) Taking the logarithm of $\tilde{H}_{k-1}$ to get the V needed for the next transformation.

The problem of finding a well-defined Lagrangian in the ultraviolet limit then reduces to seeing if V converges in the limit of infinitely many RG transformations: in the limit of

[10] See Watanabe (2000) for details.

$k \to \infty$. The convergence of V in this way corresponds to the existence of the fixed point we are looking for, as explained in Section 4.

Constructive field theory differs from other ways of implementing the RG in how well it controls the approximations that are involved. For bosonic interactions, the step of taking the logarithm of $\tilde{H}_{k-1}$ is not well-defined for certain values of ϕ. This is the large field problem. Constructive field theory deals with this by carrying out the transformation only for small fields. The steps of integrating out fluctuations in a momentum slice and taking the logarithm are carried out only for small fields. This means that we can use a cluster expansion for the former step and a Mayer expansion for the latter step. Both these expansions would not be well-controlled in the large field region. There are various methods for controlling the large field region. Because of their complexity, I can only list them here without going into the details: the domination procedure (Feldman, Magnen, Rivasseau, and Sénéor 1987), polymer systems (Pordt 1994), and using the fact that "large fields" occur with a relatively small probability (Balaban, Imbrie, and Jaffe 1984).

6. Conclusion

I have argued against a view that the RG in QFT is merely a way to get from the fundamental physics given by axiomatic QFT to macroscopic experimental predictions. Rather, the RG is also an important method in constructive field theory to figure out whether certain Lagrangians have well-defined UV limits that satisfy the axioms that we think a QFT ought to satisfy. Furthermore, the RG as employed in constructive field theory is not of questionable rigor.

The view that I criticise is one in which axiomatic QFT provides the theoretical content of QFT while the RG provides a way to get from this theoretical content to macroscopic empirical predictions. On this view, for interpretive purposes we need only focus

on axiomatic QFT. However, constructive field theory provides an important means of access to more of the theoretical content of QFT, with the RG providing a means of access even to the microscopic physics of QFT. This suggests that axiomatic QFT is at best a kind of *partial* characterization of the theoretical content of QFT. Indeed, mathematical physicists have long acknowledged that constructive QFT provides additional dynamical information that a pure axiomatic approach does not (Wightman 1976; Horuzhy 1990). If so, we should not too hastily dismiss the interpretive significance of computational methods that do not explicitly appear in the axioms of QFT, for these methods may be able to tell us if certain dynamics can occur in continuous spacetime.

REFERENCES

Abdesselam, A. (2007). A complete renormalization group trajectory between two fixed points. *Communications in Mathematical Physics 276*(3), 727–772.

Bagnuls, C. and C. Bervillier (2001). Exact renormalization group equations and the field theoretical approach to critical phenomena. *International Journal of Modern Physics A 16*(11), 1825–1845.

Bain, J. (2013). Effective field theories. In R. Batterman (Ed.), *The Oxford Handbook of Philosophy of Physics*, pp. 224–254. New York: Oxford University Press.

Balaban, T., J. Imbrie, and A. Jaffe (1984). Exact renormalization group for gauge theories. In G. 't Hooft, A. Jaffe, H. Lehmann, P. K. Mitter, I. M. Singer, and R. Stora (Eds.), *Progress in Gauge Field Theory*, pp. 79–103. New York: Plenum Press.

Benfatto, G., M. Cassandro, G. Gallavotti, F. Nicoló, E. Olivieri, E. Presutti, and E. Scacciatelli (1980). Ultraviolet stability in Euclidean scalar field theories. *Communications in Mathematical Physics 71*(2), 95–130.

Brydges, D., J. Dimock, and T. R. Hurd (1995). The short distance behavior of $(\varphi^4)_3$. *Communications in Mathematical Physics 172*(1), 143–186.

Douglas, M. (2011). Foundations of quantum field theory. In *Proceedings of Symposia in Pure Mathematics*, Volume 85, pp. 105–124.

Duncan, A. (2012). *The Conceptual Framework of Quantum Field Theory*. Oxford University Press.

Feldman, J., J. Magnen, V. Rivasseau, and R. Sénéor (1987). Construction and Borel summability of infrared ϕ^4 by a phase space expansion. *Communications in Mathematical Physics 109*(3), 437–480.

Fraser, D. (2011). How to take particle physics seriously: A further defence of axiomatic quantum field theory. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics 42*(2), 126–135.

Gawędzki, K. and A. Kupiainen (1983). Non-Gaussian fixed points of the block spin transformation. Hierarchical model approximation. *Communications in Mathematical Physics 89*(2), 191–220.

Gawędzki, K. and A. Kupiainen (1985). Exact renormalization for the Gross-Neveu model of quantum fields. *Physical Review Letters 54*, 2191–2194.

Glimm, J. and A. Jaffe (1987). *Quantum Physics: A Functional Integral Point of View* (2nd ed.). New York: Springer.

Gurau, R., V. Rivasseau, and A. Sfondrini (2014). Renormalization: an advanced overview. Online preprint. http://arxiv.org/abs/1401.5003.pdf.

Horuzhy, S. S. (1990). *Introduction to Algebraic Quantum Field Theory*. Berlin: Springer.

Huggett, N. and R. Weingard (1995). The renormalisation group and effective field theories. *Synthese 102*(1), 171–194.

Kuhlmann, M., H. Lyre, and A. Wayne (2002). Introduction. In M. Kuhlmann, H. Lyre, and A. Wayne (Eds.), *Ontological Aspects of Quantum Field Theory*, pp. 1–29. Singapore: World Scientific.

Pordt, A. (1994). On renormalization group flows and polymer algebras. In V. Rivasseau (Ed.), *Constructive Physics: Results in Field Theory, Statistical Mechanics and Condensed Matter Physics*, Volume 446 of *Lecture Notes in Physics*, pp. 51–81. Berlin: Springer.

Rivasseau, V. (1991). *From Perturbative to Constructive Renormalization*. Princeton Univ Press.

Rosten, O. J. (2012). Fundamentals of the exact renormalization group. *Physics Reports 511*(4), 177–272.

Wallace, D. (2011). Taking particle physics seriously: A critique of the algebraic approach to quantum field theory. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics 42*(2), 116–125.

Watanabe, H. (2000). Renormalization group methods in constructive field theories. *Int. J. Mod. Phys. B 14*(12n13), 1363–1398.

Wightman, A. S. (1976). Hilbert's sixth problem: mathematical treatment of the axioms of physics. In *Mathematical Developments Arising from Hilbert Problems*, pp. 147–240. Providence, Rhode Island: American Mathematical Society.

Zee, A. (2010). *Quantum Field Theory in a Nutshell* (2nd ed.). Princeton University Press.

DEPARTMENT OF HISTORY AND PHILOSOPHY OF SCIENCE, UNIVERSITY OF PITTSBURGH

E-mail address: BBL3@pitt.edu

How to Model Mechanistic Hierarchies*

Lorenzo Casini†

November 4, 2014

please do not quote or cite without permission

Abstract

Mechanisms are usually viewed as inherently hierarchical, with lower levels of a mechanism influencing, and decomposing, its higher-level behaviour. In order to adequately draw quantitative predictions from a model of a mechanism, the model needs to capture this hierarchical aspect. The recursive Bayesian network (RBN) formalism was put forward as a means to model mechanistic hierarchies (Casini et al., 2011) by decomposing *variables*. The proposal was recently criticized by Gebharter (2014) and Gebharter and Kaiser (2014), who instead propose to decompose *arrows*. In this paper, I defend the RBN account from the criticism and argue that it offers a better representation of mechanistic hierarchies than the rival account.

Contents

*To be presented at PSA 2014, Chicago, 6–8 Nov 2014, in the symposium "How Adequate Are Causal Graphs and Bayesian Networks for Modeling Biological Mechanisms?"

†Address: Department of Philosophy, University of Geneva, 5, Rue de Candolle, CH-1211 Genève 4, Switzerland. Email: lorenzo.casini@unige.ch

Introduction

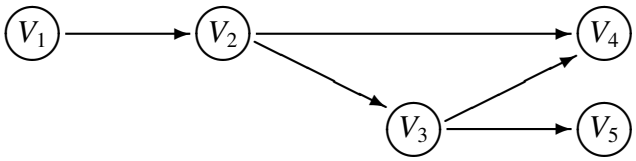
Mechanisms are usually viewed as inherently hierarchical, with lower levels of a mechanism influencing, and decomposing, its higher-level behaviour. In order to adequately draw quantitative predictions from a model of a mechanism, the model needs to capture this hierarchical aspect. The *recursive Bayesian network* (RBN) formalism was put forward as a means to model mechanistic hierarchies (Casini et al., 2011). The formalism is an extension of the Bayesian network (BN) formalism, already used to model same-level causal relations probabilistically (Pearl, 2000). In RBNs, higher-level *variables* decompose into lower-level causal BNs.

This proposal was recently criticized by Gebharter (2014) and Gebharter and Kaiser (2014), on two main grounds: descriptive adequacy—it is unclear when the formalism is applicable to real mechanisms—and conceptual adequacy—RBNs do not allow one to draw interlevel inferences for explanation and intervention. To overcome these alleged limitations, Gebharter (2014) and Gebharter and Kaiser (2014) have made the alternative proposal that decomposition involves *arrows* rather than variables. In particular, Gebharter (2014) proposes an alternative formalism, also extending the BN formalism, namely *multilevel causal models* (MLCMs). Instead, Gebharter and Kaiser (2014) make an informal proposal, which as we shall see, does not coincide with MLCMs.

Decomposing variables and decomposing arrows are two very natural options for representing mechanistic hierarchies, if one's starting point is already a probabilistic interpretation of causality. In this paper, I argue that the former option is superior to the latter. I proceed as follows. In §1 I present and illustrate RBNs and MLCMs. In §2 I argue against decomposing arrows. MLCMs lead to counterintuitive notions of mechanistic decomposition and mechanistic explanation; and Gebharter and Kaiser (2014)'s informal proposal goes only halfway towards a solution. Finally, in §3 I defend RBNs from the criticism. RBNs do allow interlevel causal explanation, via the uncoupling of interlevel causal relations into a constitutional step and a causal step. RBNs also allow reasoning about interlevel interventions; believing otherwise depends on either wrongly assuming that changes cannot transmit along the constitutional downward-directed arrows, or on demanding that the RBN formalism represent intervention variables, which the formalism is not meant to represent.

1 The two formalisms

Both RBNs and MLCMs are extensions of the BN formalism. A BN consists of a finite set $V = \{V_1, \ldots, V_n\}$ of variables, each of which takes finitely many possible values, together with a directed acyclic graph (DAG) whose nodes are the variables in V, and the probability distribution $P(V_i|Par_i)$ of each variable V_i conditional on its parents Par_i in the DAG. Here is an example:



DAG and probability function are linked by the *Markov Condition* (MC):

MC. For any $V_i \in \mathcal{V} = \{V_1, \ldots, V_n\}$, $V_i \perp\!\!\!\perp ND_i \mid Par_i$.

In words, each variable is probabilistically independent of its non-descendants, conditional on its parents. The above figure implies for instance that V_4 is independent of V_1 and V_5 conditional on V_2 and V_3. In the BN jargon, V_2 and V_3 'screen off' V_4 from V_1 and V_5. A BN determines a joint probability distribution over its nodes via $P(v_1 \cdots v_n) = \prod_{i=1}^{n} P(v_i|par_i)$ where v_i is an assignment $V_i = x$ of a value to V_i and par_i is the assignment of values to its parents induced by the assignment $v = v_1 \cdots v_n$.

In a *causally-interpreted* BN, the arrows in the DAG are interpreted as direct causal relations and the network can be used to infer the effects of interventions as well as to make probabilistic predictions (Pearl, 2000). In this case, MC is called the *Causal Markov Condition* (CMC).

1.1 Recursive Bayesian networks

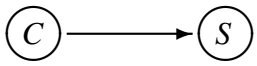
RBNs represent hierarchies by decomposing variables (Casini et al., 2011). One of the motivations behind this choice is that scientists often talk of properties at different levels that stand in a constitutive relation with one another.[1] Another

[1] Famously, Craver (2007) has proposed a criterion for identifying constitutive relations, namely the 'mutual manipulability' of higher- and lower-level properties that stand in the relation. Casini et al. (2011) refer to Craver's intuition to further motivate RBNs. Arguments against the compatibility between Craver (2007)'s account of constitution and interventionism (Woodward,

motivation—which was only implicit in (Casini et al., 2011)—is that decomposing variables has the additional advantage of making 'interlevel causation' intelligible, by uncoupling (problematic) cases of interlevel downward or upward causation into two (less-problematic) steps, a constitutional, across-level step and a causal, same-level step (Craver and Bechtel, 2007). RBNs make this idea formally precise, thereby adding an additional justification to it.

Mechanistic hierarchy is interpreted via the notion of 'recursive decomposition' of variables. An RBN is a BN defined over a finite set V of variables *whose values may themselves be RBNs*. A variable is called a *network variable* if one or more of its possible values is an RBN and a *simple variable* otherwise. A standard BN is an RBN whose variables are all simple. An RBN x that occurs as the value of a network variable in RBN y is said to be at a *lower level* than y; variables in y are the *direct superiors* of variables in x while variables in the same network are *peers*.[2] If an RBN contains no infinite descending chains—i.e., if each descending chain of networks terminates in a standard BN—then it is *well-founded*. Only well-founded RBNs are considered here.

Consider a toy RBN on $V = \{C, S\}$, where C represents whether some tissue in an organism is cancerous, taking the possible values 1 and 0, while S is survival after 5 years, taking the possible values *yes* and *no*. The corresponding BN is:



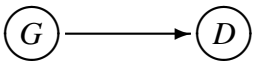
$$P(C),\ P(S|C)$$

Figure 1.1.1

2003), on which Craver's account is based, have been offered by Leuridan (2012) and Baumgartner and Gebharter (2014). Two remarks are in order. First: in the light of Gebharter and Kaiser (2014, 3.5.3)'s own endorsement of Craver (2007)'s interpretation of constitution, these arguments may be negatively relevant to *both* RBNs and Gebharter and Kaiser (2014)'s proposal. Although this issue is certainly worth considering, I do not discuss it further here. I should however point out that RBNs do not *define* constitution. They only *characterize* it, probabilistically—and *not* even in interventionist terms (cf. fn. 4). Interventions are only used to reason about interlevel causation. Second: Gebharter (2014)'s MLCM formalism does *not* interpret hierarchy in terms of constitution—let alone constitution in one specific sense. It is thus immune to this critique. However, instead of being an advantage, this threatens to undermine MLCMs' ability to represent *mechanistic* hierarchies (see §2).

[2]A variable can have several superiors. If a variable appears more than once in an RBN, the network should not imply incompatible things about it. Consistency is discussed in detail in (Williamson, 2005, §§10.4–10.5).

Suppose S is a simple variable but C is a network variable, with each of its two values denoting a lower-level (standard) BN that represents a state of the mechanism for cancer. I will ignore many of the factors, such as DNA damage response mechanisms, also responsible for cancer, and only focus on the unregulated cell growth that results from mutations in factors that control cell division, usually labelled 'growth factor', in short GF. When C is assigned value 1 we have a network c_1 representing a functioning control mechanism, with a probabilistic dependence (and a causal connection) between growth factor G and cell division D.



$P_{c_1}(G),\ P_{c_1}(D|G)$

Figure 1.1.2

On the other hand, when C is assigned value 0 we have a network c_0 representing a malfunction of the growth mechanism, with no dependence (and no causal connection) between G and D.

$P_{c_0}(G),\ P_{c_0}(D)$

Figure 1.1.3

Since these two lower-level networks are standard BNs, the RBN is well-founded and fully described by the three networks.[3]

If an RBN is to be used to model a mechanism, it is natural to interpret the arrows at the various levels of the RBN as signifying causal connections. Just as standard causally-interpreted BNs are subject to the CMC, a similar condition applies to causally-interpreted RBNs, called the *Recursive Causal Markov Condition* (RMC). Let us indicate with NID_i the set of non-inferiors-or-descendants of V_i and with $DSup_i$ the set of direct superiors of V_i. Then, RCMC says that

[3]Note that, as this example shows, an RBN may be used to represent several states of one and the same mechanism—in this case, the RBN represents a functioning state as well as a malfunctioning state. However, it need not be so used—it is also possible to build an RBN that represents just one mechanism state by having the network variable take a unique possible value.

RCMC. For any $V_i \in \mathcal{V} = \{V_1, \ldots, V_n\}$, $V_i \perp\!\!\!\perp NID_i \mid DSup_i \cup Par_i$.

In words, each variable in the RBN is independent of those variables that are neither its effects (i.e., descendants) nor its inferiors, conditional on its direct causes (i.e., parents) and its direct superiors. RCMC adds to CMC the condition that variables at different levels also stand in relations that fulfil a MC, namely variables at any level are probabilistically independent of non-inferiors or peers given their direct superiors. Intuitively, if one knows the value of C, knowledge of the value of constituent variables G or D doesn't add anything to one's ability to infer to, say, the causes of C (here, none) or to the effects of C (here, S). Since the screening off that holds in virtue of RMC depends on constitutional rather than causal facts, not all dependencies identified by the RCMC can be causally interpreted.

Notice that, while some authors treat CMC as a necessary truth, others argue against its universal validity (see, e.g., Williamson, 2005). Here a similar stance is adopted with respect to RCMC. RCMC is a *modelling assumptions* in need of testing or justification, rather than as a necessary truth. From this, it follows that whether or not the formalism allows one to adequately represent a mechanism is an empirical matter, rather than a matter of stipulation. For instance, whether or not C adequately screens off S from G and D depends, among other things, on the assumption that G and D affect S only via C. If this is not true, because S or G participate in other mechanisms for S, RCMC is violated. Recovering RCMC would then require including other network variables that cause S, and that decompose into, among other variables, G and/or D.

Inference in RBNs proceeds via a formal device called a *flattening*. Let $\mathcal{V} = \{V_1, \ldots, V_m\}$ $(m \geq n)$ be the set of variables of an RBN closed under the inferiority relation: i.e., $\mathcal{V}$ contains the variables in V, their direct inferiors, their direct inferiors, and so on. Let $\mathcal{N} = \{V_{j_1}, \ldots, V_{j_k}\} \subseteq \mathcal{V}$ be the network variables in $\mathcal{V}$. For each assignment $n = v_{j_1}, \ldots, v_{j_k}$ of values to the network variables we can construct a standard BN, the *flattening* of the RBN with respect to n, denoted by $n^{\downarrow}$, by taking as nodes the simple variables in $\mathcal{V}$ plus the assignments $v_{j_1}, \ldots, v_{j_k}$ to the network variables, and including an arrow from one variable to another if the former is a parent or direct superior of the latter in the original RBN. The conditional probability distributions are constrained by those in the original RBN—in the RBN where V_{j_i} is the direct superior of V_i, $P(V_i|Par_i \cup DSup_i) = P_{v_{j_i}}(V_i|Par_i)$. Notice that MC holds in the flattening because the RCMC holds in the RBN. Only, since the arrows in the flattening that link variables to their direct inferiors are constitutional, CMC is not satisfied.[4]

[4] It should now be clear that the role of RCMC—and of RBNs more generally (see fn. 1)—is

The flattenings suffice to determine a joint distribution over the variables in $\mathcal{V}$ via $P(v_1 \cdots v_m) = \prod_{i=1}^{m} P(v_i|par_i dsup_i)$ where the probabilities on the right-hand side are determined by a flattening induced by $v_1 \cdots v_m$.[5] In the cancer example, for assignment c_1 of network variable C we have the flattening $c_1^{\downarrow}$:

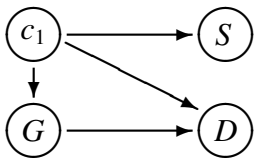


Figure 1.1.4

with probability distributions $P(c_1) = 1$ and $P(S|c_1)$ determined by the top level of the RBN, and with $P(d_1|g_1c_1) = P_{c_1}(d_1|g_1)$ determined by the lower level (similarly for g_0 and d_0). The flattening with respect to assignment c_0 is $c_0^{\downarrow}$:

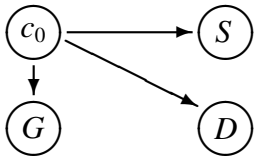


Figure 1.1.5

Again, $P(d_1|c_0) = P_{c_0}(d_1)$ etc. In each case the required conditional distributions are determined by the distributions given in the original RBN.

Having determined a joint distribution, the causally-interpreted RBN may, in just the same way as can a standard causal BN, be used to draw quantitative inferences for explanation and intervention, inferences that may involve variables at the same level as well as—so we claimed in (Casini et al., 2011, §2)—*across* levels.

not to define constitutional relations. With respect to the flattening, the choice of calling some arrows 'causal' and other arrows 'constitutional' is not dictated by MC. Any use of RCMC to find out what does (not) constitute what *presupposes* a prior distinction between the variables at the different levels. Yet, given the distinction between the levels, RCMC does characterize constitutional relations in terms of certain probabilistic dependencies and independencies.

[5] $P_{v_{j_l}}(V_i \mid Par_i)$ may be obtained from observed frequencies in a dataset. Instead, $P(V_i \mid Par_i DSup_i)$ can be obtained in either of two main ways. Either one determines the corresponding observed frequencies from the original dataset, or one selects from all functions that satisfy the probabilistic constraints imposed by the RBN the function Q with maximum entropy (Williamson, 2010), and sets $P(V_i \mid Par_i DSup_i) = Q(V_i \mid Par_i DSup_i)$.

1.2 Multilevel causal models

According to Gebharter (2014), RBNs fail to allow interlevel causal inferences, due to the lack of an explicit representation of interlevel causal arrows, over which causal influence propagates. These objections, I maintain, are based on the (mis)interpretation of RBNs. I postpone this discussion to §3.

Gebharter's proposed formalism purports to remedy these alleged deficiencies by decomposing causal *arrows* rather than variables. More precisely, mechanistic hierarchy has for him to do with 'marginalizing out' variables when moving from a lower-level graph to a higher-level graph.

Let us indicate a causal model as $\langle V, E, P\rangle$, where $\langle V, E\rangle$ is a DAG, defined over a variable set V and a set of edges E among them, and P an associated probability distribution. Let $X \leftrightarrow Y$ indicate that two variables X and Y are effects of a latent common cause, i.e., a cause of X and Y not represented within the graph of some variable set V. Also, let us indicate with $P^* \uparrow V$ the 'restriction' of the probability distribution P^* to variable set V. The restriction of a lower-level causal model $\langle V^*, E^*, P^*\rangle$ to a higher-level causal model $\langle V, E, P\rangle$ is so defined (2014, 147):

Restriction. $\langle V, E, P\rangle$ is a restriction of $\langle V^*, E^*, P^*\rangle$ if and only if

- **a** $V \subset V^*$, and
- **b** $P^* \uparrow V = P$, and
- **c** for all $X, Y \in V$:
  - **c.1** if there is a directed path from X to Y in $\langle V^*, E^*\rangle$ and no vertex on this path different from X and Y is in V, then $X \rightarrow Y$ is in $\langle V, E\rangle$, and
  - **c.2** if X and Y are connected by a common cause path π in $\langle V^*, E^*\rangle$ or by a path π free of colliders containing a bidirected edge in $\langle V^*, E^*\rangle$, and no vertex on this path π different from X and Y is in V, then $X \leftrightarrow Y$ is in $\langle V, E\rangle$, and
- **d** no path not implied by **c** is in $\langle V, E\rangle$.

That is, the lower-level structure $\langle V^*, E^*, P^*\rangle$ represents the mechanism for the higher-level structure $\langle V, E, P\rangle$ iff $\langle V, E, P\rangle$ is the restriction of $\langle V^*, E^*, P^*\rangle$ uniquely determined when V^* is restricted to V. The restriction is such that all and only the directed paths and common cause paths in $\langle V^*, E^*\rangle$ are preserved by $\langle V, E\rangle$, and the probabilistic information of P^* is consistent with P upon marginalizing out variables in $\{V^* \setminus V\}$.

A "multi-level causal model" (MLCM) is then so defined (2014, 148):

MLCM. $\langle M_1 = \langle V_1, E_1, P_1\rangle, \ldots, M_n = \langle V_n, E_n, P_n\rangle\rangle$ is a multi-level causal model if and only if

- **a** $M_1, \ldots, M_n$ are causal models, and
- **b** every M_i with $1 < i \leq n$ is a restriction of M_1, and
- **c** M_1 satisfies CMC.

That is, a MLCM is an ordered set of causal models $\langle M_1 = \langle V_1, E_1, P_1\rangle, \ldots, M_n = \langle V_n, E_n, P_n\rangle\rangle$, where the bottom-level, unrestricted causal model M_1 satisfies CMC. (Instead, higher-level models may or may not satisfy CMC.) Each causal model in the MLCM, for Gebharter, represents a mechanism.

The information on the hierarchical relations among the nested mechanisms in the MLCM is contained in a "level graph", which is so defined (2014, 149):

Level graph. A graph $G = \langle V, E\rangle$ is called an MLCM $\langle M_1 = \langle V_1, E_1, P_1\rangle, \ldots, M_n = \langle V_n, E_n, P_n\rangle\rangle$'s *level graph* if and only if

- **a** $V = \{M_1, \ldots, M_n\}$, and
- **b** for all $M_i = \langle V_i, E_i, P_i\rangle$ and $Mj = \langle V_j, E_j, P_j\rangle$ in V: $M_i \rightarrow M_j$ is in G if and only if $V_i \subset V_j$ and there is no $M_k = \langle V_k, E_k, P_k\rangle$ in V such that $V_i \subset V_k \subset V_j$ holds.

A level graph $G = \langle V, E\rangle$ is constructed from a MLCM by adding dashed (non-causal) arrows between any two models M_i and M_j, $M_i \rightarrow M_j$, if and only if V_i is the largest proper subset of V_j in MLCM, so that M_i is, so to say, the smallest restriction of M_j. Here is an example of level graph from (Gebharter, 2014, 150):

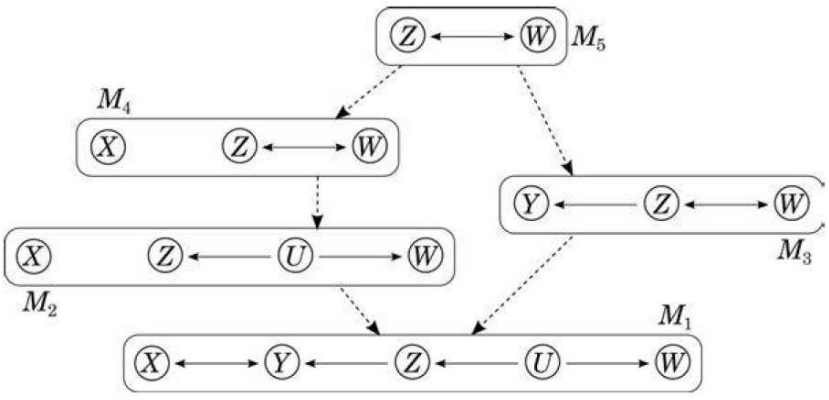


Figure 1.2.1

Notice that the ordering among graphs is not strict, so there may be pairs of graphs (e.g.: M_2 and M_3; M_4 and M_3) that do not stand in a restriction relation.

Below is a more concrete illustration from (Gebharter, 2014, 151), the representation of a water dispenser mechanism, on two levels,

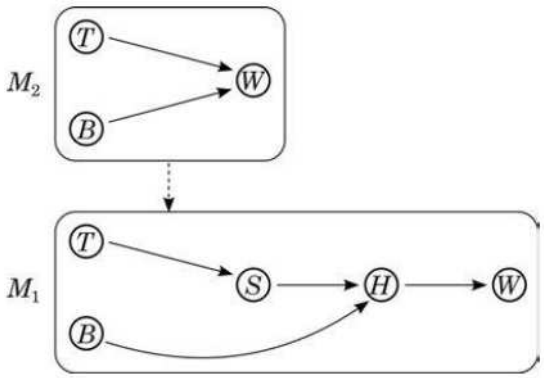


Figure 1.2.2

such that M_1 contains the following direct causal relations: the room temperature T activates (and is measured by) a sensor S; S, together with the status of a tempering button, B, cause the heater to be on or off, H; H in turn causes the temperature of the water dispensed, W.[6]

2 Criticism of MLCMs

It is debatable whether hierarchies, as represented by the level graphs in figures 1.2.1 and 1.2.2, are *mechanistic*—whether they represent mechanistic decompositions, and grant mechanistic explanations.

[6]Gebharter contrasts the virtues of this MLCM with an RBN of the 'same' mechanism (2014, 142-3). However, this is somewhat misleading. Gebharter's RBN is defined over a larger variable set, which includes a network binary variable D, superior to S and H, caused by T and B, and causing W. It is obvious that his RBN cannot represent the *same* mechanism as his MLCM. On the assumption that the RBN is faithful, it should be possible to order the RBN's flattening (Gebharter, 2014, 144), call it M_0, as *prior* with respect to M_1—since M_1's variable set V_1 is $\{V_0 \setminus D\}$. However, M_1 is incompatible with the restriction of M_0 obtained by marginalizing out D, call this M_{1^*}. (M_{1^*} would contain $S \leftrightarrow H$, $S \leftrightarrow W$, $H \leftrightarrow W$ and $B \rightarrow S$. Instead, M_1 contains $S \rightarrow H$, $H \rightarrow W$ and $B \rightarrow H$.) Thus, rather than one model being a correct representation and the other being a wrong representation of one and the same mechanism, the two models represent different mechanisms, and are thus are not directly comparable. In the following, I shall defend RBNs with reference to the toy model introduced in §1.1.

First, it is not clear if MLCMs adequately represent mechanistic decompositions. High-level causal models in a MLCM, for instance models M_2 in figure 1.2.1, are just more coarse-grain representations of one and the same mechanism, viz. M_1, such that some of the information in M_1 is *missing* at the higher level, as the term 'restriction' suggests. Is, for instance, $T \rightarrow S \rightarrow H \rightarrow W$ a mechanistic decomposition of $T \rightarrow W$, although *entities* and *properties* involved are the same at both levels, and only some *activities* (or relations) are different? Perhaps this counts as a different, equally legitimate, notion of decomposition, call it *decomposition**. The question is: How intuitive is *decomposition**?

Second, it is not clear if MLCMs adequately represent mechanistic explanations. One may concede that there is a legitimate sense in which one explains the relation between, say, the room temperature T and the water temperature W by blowing up the process from the former to the latter and uncovering the mediating role of the sensor S and the heater H. However, this sort of explanation is different from the equally legitimate explanation whereby one redescribes the cancer mechanism C in figure 1.1.1 into more fine-grain terms, and uncovers the role of damage G and response D. G and D have an obvious mechanistic role. Instead, S and H seem to have an *etiological* role. Perhaps S and H still explain mechanistically, according to some different notion of mechanistic explanation, call it *explanation**. But just how intuitive is *explanation**?

The counterintuitive nature of *decomposition** and *explanation** is made more explicit by a careful scrutiny of the level graph in figure 1.2.1. To begin with, consider the 'decompositions' that correspond to restricting (i) V_1 to V_2, (ii) V_1 to V_3, and (iii) V_3 to V_5. In all such cases, instead of opening a black box (as is common in mechanistic explanation), one 'creates' a box, and does not, strictly speaking, decompose anything. Let us consider (i). Here the decomposition is 'filling a blank': the absence of probabilistic and causal dependencies among variables is explained by direct causation, a hidden common cause structure, or combinations thereof that involve new variables, too. The absence of probabilistic and causal dependencies between X and Z in M_2 is explained by the structure $X \leftrightarrow Y \leftarrow Z$ in M_1 (more on this alleged case of 'explanation' below). Since there is no arrow between X and Z in M_2, and since mechanisms require causal dependencies, what mechanism is $X \leftrightarrow Y \leftarrow Z$ in M_1 a decomposition of? Next, consider cases (ii) and (iii). Here the decomposition is in fact 'adding stuff'. For instance, $Z \leftrightarrow W$ in M_5 is 'decomposed' into $Y \leftarrow Z \leftrightarrow W$ in M_3. But in what sense is a lower-level mechanism that includes an isolated effect not included in the higher level a decomposition of the higher level mechanism?

Relatedly, to some of the represented restrictions do not seem to correspond

'explanations' either. Consider the restriction of M_4 to M_5. Here, the common cause structure $Z \leftrightarrow W$ is 'explained' by the absence of probabilistic or causal dependence between Z and a new variable X, which is apparently disconnected from whatever mechanism is responsible for $Z \leftrightarrow W$. An even more striking case of lack of explanation is the 'decomposition' of X and Z in M_2 into $X \leftrightarrow Y \leftarrow Z$ in M_1. A first and more obvious issue, which is clearly non-intentional, is that the presence of a bidirected arrow in M_1 violates condition **c** of a MLCM, namely that M_1 satisfies CMC.[7] Still, even if condition **c** is satisfied, the more general problem remains that, if decompositions are to explain, this sort of decomposition should not be allowed at *any* level. Intuitively, hidden common cause structures such as $X \leftrightarrow Y$ are just that, *hidden*, and thus non-explanatory. They add a mystery rather than remove it. A—drastic—solution that immediately comes to mind is to forbid bidirected arrows at any level. This would entail, however, that restrictions that marginalize out common causes are disallowed, too, which is undesirable because—if one buys into the MLCM framework—the corresponding decompositions would seem (more) explanatory. One may of course impose further conditions that distinguish good from bad restrictions. However, it is not obvious how one should proceed in a non *ad hoc* way, in the absence of clear intuitions on the explanatory value of bidirected causal arrows.

The above reasons lead to scepticism about the formalism's capacity to represent mechanistic decompositions and explanations. Such worries are in part, but not fully, mitigated by the (orthogonal) suggestion in (Gebharter and Kaiser, 2014) that levels be ontologically distinct and the requirement that hierarchical relations are (partly) defined by constitutional part-whole relations.

> In our approach one can generate a hierarchic causal model by replacing such a causal arrow [between two variables X and Y] by another causal structure. This causal structure should be on a lower ontological level than X and Y, it should contain at least one constitutively relevant part of X and at least one of Y, and there should be at least one causal path going from the former to the latter at the micro- level. (Gebharter and Kaiser, 2014, §3.6)

In the paper, Gebharter and Kaiser focus on modelling this sort of hierarchical relation with reference to the inhibitory feedback mechanism for the regulation of

[7] Gebharter himself emphasizes that "the graph of a causal model that contains bidirected arrows no longer determines the Markov factorization [...]." (2014, 146, fn. 8).

the biosynthesis of fatty acids in Brassica napus. The mechanism may be represented as follows (see figure 1.2).

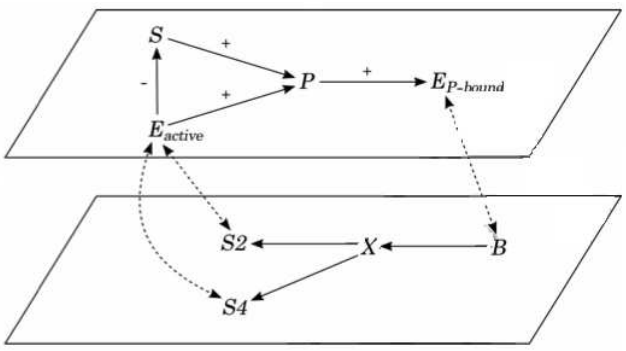


Figure 2

The product of a reaction pathway, in this case the 18:1-acyl carrier protein (P) acts as a feedback signal, which inhibits an enzyme earlier in the pathway, in this case the plastidic acetyl-CoA carboxylase (ACCase), whose operation promotes the production of P itself via the transformation of the substrate acetyl-CoA (S). ACCase has two relevant properties: it is a (positive) cause of the concentration of P (E_{active}); and it is (in its P-bound state) an effect of the concentration of P ($E_{P-bound}$). $E_{P-bound}$ is in turn a (negative) cause of E_{active} (because P-bound ACCase becomes inactive) and so on and so forth, in a cycle. In addition, E_{active} is also a negative cause on $E_{P-bound}$, which is represented by a negative influence on S. Between the binding of P to E and the inactivation of E a lower-level mechanism takes place, namely the conformational change of the substrate binding site. The binding B between functional groups of 18:1-acyl and the effector interaction site of the enzyme causes an allosteric switch X, which in turn brings about changes at sites $S2$ and $S4$ of the enzyme ACCase. This, then, prevents the substrate from being able to bind to the enzyme.[8]

It is now demanded that the levels be ontologically distinct, partly by way of decomposing properties, rather than just the relation $E_{P-bound} \rightarrow E_{active}$, as follows: B is a property of a part contained in the whole that has the property $E_{P-bound}$; and $S4$ and $S2$ are properties of parts contained in the whole that has the property E_{active}. Between parts and wholes there are relations of constitutive

[8]To get a causal model, Gebharter and Kaiser propose that the causal graph in figure 1.2 be associated with a probability distribution over a variable set that unrolls the cycle, so as to get a *dynamic* causal graph. This way of treating cycles is similar to the one adopted in the RBN approach (Clarke et al., 2014), with the notable difference that MC is not satisfied here (see below).

rather than causal relevance, in the sense of Craver (2007): a change in a part results in a change in the whole, and vice versa. More precisely, constitutive relations are represented by dashed two-headed arrows that stand at either side of the decomposition relation. As a result, decomposing arrows should apparently explain both causally *and* constitutionally.

Gebharter and Kaiser require that a causal arrow $X \rightarrow Y$ is decomposed by a lower-level causal structure only if it contains at least one constitutively relevant part of X and at least one of Y, and there is at least one causal path going from the former to the latter at the microlevel (2014). This eliminates two counterintuitive features of MLCMs, namely that mechanistic decompositions may 'fill blanks' (there must be a higher-level relation to begin with) and 'add stuff' (there must be at least one lower-level causal path). Still, two questions arise, related to the interpretation of the dashed bidirected arrows.

First, is this interpretation of mechanistic hierarchy compatible with MLCMs? As Gebharter and Kaiser notice, "since the two-headed dashed arrows in our hierarchic dynamic CM transport the influences of interventions in both directions, CMC does not hold in such models". Since M_1 would contain bidirected arrows, too, it would not satisfy CMC. This entails that the Brassica napus mechanism cannot be represented by the MLCM formalism.

Second, does the causal model in (Gebharter and Kaiser, 2014, §3.5) offer an adequate formal representation of a mechanistic hierarchy, alternative to MCLMs? I think that a positive answer would require that constitutional relations be ascribed distinctive *formal* properties. Although constitutional relations are characterized informally by part-whole relations, they don't come with distinctive probabilistic features, as one would expect from a probabilistic representation of mechanistic hierarchies. In contrast, RBNs do offer a probabilistic characterization of constitution: properties at different levels that stand in a constitutional relation relate to other properties as described by RCMC.[9]

3 Defense of RBNs

Still, the shortcomings of MLCMs would be a small consolation for the RBN advocate, if RBNs did not survive the objections raised by Gebharter (2014) and Gebharter and Kaiser (2014). In this section I will consider, and try to rebut, such objections one by one. RBNs interpret mechanistic hierarchy via the operation of

[9]To reiterate a point already made in fn. 4, RCMC does not itself distinguish the levels, and thus it cannot be used to define constitution. Still, it does characterize it.

'recursive decomposition', which in turn depends on RCMC. Two kinds of objections are raised against RCMC. First, about empirical adequacy: it is unclear when RCMC holds, so it is unclear if the formalism is applicable to real mechanisms. Second, about conceptual adequacy: RCMC prevents RBNs from being useful for interlevel reasoning for explanation and intervention. Let us begin with the first objection:

> it is neither obvious that RCMC holds in general, nor is it clear how one could distinguish cases in which it holds from cases in which it does not. (Gebharter and Kaiser, 2014, §3.5.3)

Agreed, RCMC may not hold in general. But Casini et al. (2011) don't claim that it does. When *does* it hold, then? What RCMC adds to CMC, which is not called into question here, is RMC. RMC has to do with the (in)dependencies among variables at different levels. In the cancer example, RMC depends on C screening off G and D from S.

Gebharter and Kaiser then argue that the RBN approach would be unable to adequately model the $E_{P-bound} \rightarrow E_{active}$ mechanistic decomposition:

> it is not clear how the submechanism represented by $E_{P-bound} \rightarrow E_{active}$ could be analyzed in Casini et al.'s (2011) approach. They would need to add a network variable N between $E_{P-bound} \rightarrow E_{active}$ ($E_{P-bound} \rightarrow N \rightarrow E_{active}$). But then and because there is no intermediate (macro-level) cause N between $E_{P-bound}$ and E_{active}, it is unclear what this network variable N should represent at the mechanism's macro-level. (Gebharter and Kaiser, 2014, §3.5.3)

I do not dispute that there may be cases where it is hard or implausible to find network variables that stand for lower-level causal structures. However, this is an empirical problem, and not necessarily a problem with the formalism. RBNs are meant to represent a natural decomposition strategy of functional properties into structural properties. The structural properties may be then regarded as functional with respect to other structural properties, and so on and so forth. When does a network variable N exist? This depends on identifying properties at different levels, which in turn depends on a meaningful distinction between the levels.

I propose a few conditions for distinguishing between variables in a constitutional relation.[10] First, between the whole and its parts are mereological relations, such that properties of the whole can be explained by their probabilistic

[10] I don't claim that the list is exhaustive or that each of the listed conditions is necessary.

dependence on the structure of (causal relations among) its parts' properties. Second, properties at the different levels have different explanatory roles, such that they typically enter into causal explanations involving different sets of properties. Third, there is a difference in epistemic conditions, such that the way one observes, or intervenes on, a variable at some higher level does not coincide with the the way one observes, and intervenes on, one of its constituting variables at the lower level.[11] When a distinction between variables informed by the above conditions is possible, the distinction between the levels seems legitimate.[12]

A network variable N exists insofar as the lower-level BN is the decomposition of one functional property, which, according to the aforementioned criteria, corresponds to a *whole*'s property that has its own *explanatory role* and *epistemic autonomy*. These conditions seem satisfied by many descriptions of mechanisms in science. For instance, tissues are made of cells. Scientists talk of the cancerous state of a tissue as having an explanatory role with respect to survival. One may observe the state of a tissue or change it, for instance by replacing the whole tissue. One may use this knowledge to then infer to the probability of survival. This does not require knowing, or (surgically) intervening on, the state of GF. [13]

Finally, let us come to the objection that RBNs do not support interlevel reasoning for explanation and for prediction of the results of interventions:

> [Casini et al.'s] approach does (i) not allow for a graphical representation of how a mechanism's macro variables are causally connected to the mechanism's causal micro structure, which is essential when it comes to causal explanation, and it (ii) leads to the fatal consequence that a mechanism's macro variables' values cannot be changed by any intervention on the mechanism's micro structure whatsoever [...] (Gebharter, 2014, 139)

Explanation first. Since there are no arrows between variable at different levels screened off by network variables, Gebharter claims that it is unclear over which causal paths probabilistic influence propagates between such higher- and lower-level variables (cf. 2014, 143-4). I reply that it is true, there are no such arrows.

[11] Baumgartner and Gebharter (2014) develop this intuition into a 'fat-handedness' criterion for constitution. (Ironically, there an argument is proposed to defend an interpretation of mechanistic hierarchy based on decomposing variables rather than arrows.)

[12] The conditions only provide a useful heuristics. They do not belong to the RBN formalism. Still, RBNs give a probabilistic characterization of constitution, thanks to RCMC (cf. fn. 4).

[13] For more realistic examples, see (Casini et al., 2011), (Clarke et al., 2014) and (Casini, 2014).

But this is because, by assumption, screened off variables influence each other, if at all, only *via* the network variables. So, when RCMC is satisfied, the probabilistic influence propagates *constitutionally* (rather than causally) across the dashed arrows in the flattenings, and causally across the same-level solid arrows.

Let us now consider the second objection. With reference to the example in figures 1.1.4 and 1.1.5, I claimed that one may, for instance, reason about the result of a lower-level intervention on D on the probability of the higher-level variable S. Given the observed value of $P(s_1)$, calculated as

$$P(s_1) = P(c_0)P(s_1|c_0) + P(c_1)P(s_1|c_1),$$

one may ask: What is the effect of setting $D = d_1$ on the probability of observing $S = s_1$? To answer, one calculates as follows. First, one removes the arrow $G \rightarrow D$ from c_1, so that both flattenings have the same structure below.

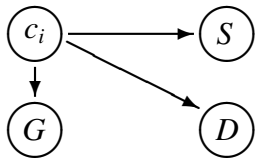


Figure 3.1

Then, one calculates $P(s_1||d_1) = P(s_1d_1)/P(d_1)$, where:

$$P(s_1d_1) = P(c_0s_1d_1) + P(c_1s_1d_1) = P(c_0)P(s_1|c_0)P_{c_0}(d_1) + P(c_1)P(s_1|c_1)P_{c_1}(d_1);$$

$$P(d_1) = P(c_0)P_{c_0}(d_1) + P(c_1)P_{c_1}(d_1).$$

Gebharter objects that "according to the RBN approach, intervening on a mechanism's microvariables does not have any probabilistic influence on any one of the macrovariables whatsoever" (2014, 145) *because* if one were to use an intervention variable I to intervene on a lower-level variable, the intervention "would—and this can directly be read off the BN's associated graph's topology [...]—not have any probabilistic influence on any macrovariable at all" (*ibid.*). In the cancer example: an intervention I_R on R would not have any effect on S. There is either one of the following problems with this objection.

First, it is true that c_i screens off D from S, and thus there is no $D \rightarrow S$ causal arrow. However, concluding that interventions on R can make no difference to S would be wrong. The lack of *causal* connections in the flattening does not block

changes along *constitutional* arrows. It is important to stress that, although the dashed arrows point downwards in the flattening, this is due to technical reasons only, having to do with the condition for MC to hold across levels. Still, one may use the downward-pointing arrows to reason—constitutionally—in both directions. Here, changing D makes a constitutional difference to C, which makes a causal difference to S. The overall difference is calculated with the RBN.

Second, there may be a more basic interpretive problem regarding how interventions are represented in RBNs. True, RCMC says that S is independent of any variable that is not an effect or an inferior (here, none), conditional on its direct causes (here, C) and direct superiors (here, none). But notice that RCMC is assumed to hold true of variables in $\mathcal{V} = \{M, S, G, D\}$, and *not* of such an expanded $\mathcal{V}^+ = \{M, S, G, D, I_D\}$. The reason for this is not *ad hoc*. RBNs are meant to represent decompositions of (properties of) wholes into (properties of) their parts. They are *not* meant to represent parts that do not belong to any whole—which is what I_D is. The graph topology cannot represent such parts. As a result, one can*not* read off the graph topology that such interventions variables have no effect. More generally, in an RBN, everything one gets at lower levels must be the result of (recursively) decomposing the top level.

This should not be seen as a limitation, but as a means to achieve some end. In the RBN formalism one cannot represent interventions *as variables*—unless the variables describe properties of either the top level mechanism or of *sub*mechanisms at some lower level, obtained by way of (recursive) decompositions. But this would mean that the intervention is *not external* to the mechanism, contrary to the original intention. One can, instead, straightforwardly represent interventions *as (new) values* of either top-level variables or lower-level variables into which network variables (recursively) decompose. The two ways correspond to two well-known ways for representing interventions. Woodward (2003)'interventionist semantics, which represents interventions as variables, is an example of the former. Pearl (2000)'s *do*-calculus, which represents interventions as values of variables, is an an example of the latter. Although both representations are legitimate, only the latter is suitable to the task for which RBNs were developed, namely to represent mechanistic decompositions.

Conclusion

Decomposing variables and decomposing arrows are two very natural options for representing mechanistic hierarchies by means of BNs. These two options have

been made precise by two formalisms, RBNs and MLCMs. I argued that RBNs are better than MLCMs at analysing mechanistic hierarchies and interpreting the interlevel reasoning that depends on them. Still, one might think that the two formalisms are not in competition against one another. Perhaps RBNs and MLCMs represent two different ways in which mechanistic decompositions cab obtain? Since 'marginalizing out' and 'recursively decomposing' are very different notions, I want to caution against interpreting the two formalisms as two species of the same genus. Having said this, I do not exclude that there is a sound way to formalize the intuition in (Gebharter and Kaiser, 2014), and thus develop an alternative analysis of mechanistic hierarchy with respect to RBNs. In that case, it would be interesting to see how this alternative relates to RBNs.

Acknowledgments I wish to thank the participants to the *Biological Interest Group* of the Lake Geneva, where a prior version of this paper was discussed on 28 October 2014. I am also grateful to Jon Williamson for helpful discussions on this topic.

References

Baumgartner, M. and Gebharter, A. (2014). Constitutive Relevance, Mutual Manipulability, and Fat-handedness. *British Journal for the Philosophy of Science*. forthcoming.

Casini, L. (2014). Failures of Modularity and Recursive Bayesian Networks. Unpublished.

Casini, L., Illari, P. M., Russo, F., and Williamson, J. (2011). Models for Prediction, Explanation and Control: Recursive Bayesian Networks. *THEORIA*, 70:5–33.

Clarke, B., Leuridan, B., and Williamson, J. (2014). Modeling Mechanisms with Causal Cycles. *Synthese*, 191:1651–1681.

Craver, C. and Bechtel, W. (2007). Top-down Causation Without Top-down Causes. *Biology and Philosophy*, 22:547–563.

Craver, C. F. (2007). *Explaining the Brain*. Oxford: Oxford University Press.

Gebharter, A. (2014). A Formal Framework for Representing Mechanisms? *Philosophy of Science*, 81(1):138–153.

Gebharter, A. and Kaiser, M. I. (2014). Causal Graphs and Biological Mechanisms. In Kaiser, M. I., Scholz, O., Plenge, D., and Hüttemann, A., editors, *Explanation in the Special Sciences: The Case of Biology and History*, pages 55–85. Dordrecht: Springer.

Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *British Journal for the Philosophy of Science*, 63:399–427.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Williamson, J. (2005). *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford: Oxford University Press.

How-Possibly Explanations in Quantum Computer Science

Michael E. Cuffaro

Ludwig-Maximilians-Universität München,

Munich Center for Mathematical Philosophy

Abstract

A primary goal of quantum computer science is to find an explanation for the fact that quantum computers are more powerful than classical computers. In this paper I argue that to answer this question is to compare algorithmic processes of various kinds, and in so doing to describe the possibility spaces associated with these processes. By doing this we explain how it is possible for one process to outperform its rival. Further, in this and similar examples little is gained in subsequently asking a how-actually question. Once one has explained how-possibly there is little left to do.

Word count: 4,927.

1 Introduction

There is a distinction that is sometimes made in the scholarship on scientific explanation between explaining *why* and explaining *how-possibly*. In the ontic context, where the explanations one gives aim at describing salient features of actual physical systems, the former is sometimes also called *how-actually* explanation.[1] That how-actually explanation actually explains is uncontroversial; however it is less clear just what if any explanatory merit there is in explaining how some event possibly came about. Partly for this reason, the literature on how-possibly explanation is comparatively sparse, and the few who have commented on the topic are of varying opinion with regard to its virtues. While some view how-possibly explanation as genuinely explanatory, others have argued that how-possibly 'explanation' is better thought of as a mere heuristic device and not as constituting genuine explanation at all. Still others have thought of how-possibly explanation as a kind of incomplete how-actually explanation—a stepping stone on the way to the how-actually explanation that one ultimately seeks.

Below I will consider a question which I will argue sheds light on this issue. It is drawn from the science of *quantum computation*. Quantum computation is a fruitful merger of the fields of physics and computer science, and one of the goals of this science is to determine the source of the power of quantum computers; i.e., to search for the explanation of the fact that quantum computers can in general (and sometimes dramatically) outperform classical computers. What I will argue is the following: to answer this question is to compare algorithmic processes of various kinds, and in so doing to describe the possibility spaces

[1] For more on the distinction between ontic, epistemic, and modal forms of explanation, see Salmon (1984).

associated with these processes. By doing this we explain how it is possible for one process to outperform its rival. Further, and importantly, in examples like these little if anything is gained in subsequently asking a how-actually question. Once one has answered the how-possibly question there is little left to do.

I will close by suggesting that the search for the explanation of the power of quantum computation is just one example of a species of how-possibly question that is likely to be found in many other sciences as well.

2 How-possibly Explanation

The first mention of how-possibly explanation is likely that of Dray (1957). Dray's primary goal in that book is to assess the adequacy of the 'covering law model' of explanation for characterising historical explanation. The model is so-called because it involves the subsumption of a particular set of initial conditions under a law or a set of laws (Hempel and Oppenheim, 1948). Dray's verdict is that the covering law model fails to capture many interesting senses of historical explanation. One of the ways in which it is inadequate, according to Dray, is that the covering law model insists that any explanation of a given fact must show why, necessarily, that fact had to occur, since the statement of the fact to be explained must be deductively entailed by the statements of the relevant laws and initial conditions.[2] Dray insists, however, that not all historical explanations are why-necessarily explanations.

> An announcer broadcasting a baseball game from Victoria B.C., said: "It's a long

[2]This is the case for Hempel and Oppenheim's *Deductive-Nomological* (D-N) model of explanation. Statistical explanations require only inductive support (Hempel, 1965). For our purposes we need only detain ourselves with the former.

> fly ball to centre field, and it's going to hit high up on the fence. The centre fielder's back, he's under it, he's caught it, and the batter is out." Listeners who knew the fence was twenty feet high couldn't figure out how the fielder caught the ball. Spectators could have given the unlikely explanation. At the rear of centre field was a high platform for the scorekeeper. The centre fielder ran up the ladder and caught the ball twenty feet above the ground (from *Maclean's Magazine*, as cited in Dray 1957, 158).

What is explained here, for Dray, is not exactly why the ball landed in the centre fielder's glove. Rather, what is dispelled is the initial puzzlement on the part of the listener upon hearing about the catch. This puzzlement is removed once she is told of the scorekeeper's ladder, for the ladder explains how the catch was possible: it opens up a range of possibilities that would not have been present without it. Of course one can still ask: "why, exactly, did the ball land in his glove?" However to do so, for Dray, is to ask a logically *different* question. The how-possibly question is answered once we have been told about the ladder.

Hempel and Oppenheim's covering law model, with its exclusive emphasis on why-necessarily questions, was, for many years, the 'received view' on scientific explanation. But although similar conceptions continue to be defended, there is no longer a near consensus, and indeed many have taken a pluralistic attitude (or at any rate remained agnostic), on the question of whether an all-encompassing model of scientific explanation exists.[3] Despite this, how-possibly explanation has received comparatively little attention (as compared with, say, causal explanation). But it has received some. There are those, for instance, who reject outright the very idea of how-possibly explanation—Reiner (1993, 68) goes so far as to call

[3]See, for instance, Woodward (2003), who productively focuses his energies on explicating one particular type of explanation.

its promotion and proliferation a "sociological risk"—however the majority of the debate surrounding how-possibly explanation centres around the sense and extent to which it (in Dray's or perhaps some other formulation[4]) is explanatory.

Thus even Hempel grants to Dray that there is some sense in which a how-possibly account explains. Nevertheless, he argues, upon hearing it the questioner will invariably desire to be told why the event necessarily occurred if he is to be fully satisfied (Hempel, 1965, 429). For Hempel, the role of how-possibly explanation is primarily pragmatic: it motivates the questioner to ask a further why-necessarily question. For Resnik (1991, 143), on the other hand, how-possibly explanation and how-actually explanation are of the same kind, and differ only in the degree to which they are empirically supported. That is, a how-possibly explanation is a how-actually explanation that enjoys no more than speculative supporting evidence, yet nevertheless displays other explanatory virtues like fruitfulness. Salmon's (1989, 137) conception is similar.

Forber (2010), in contrast, views how-possibly and how-actually explanation as different in kind. What Resnik refers to as how-possibly explanation is, for Forber, no more than an

[4]Dray's account is sometimes thought to rely overmuch on psychology: for Dray, recall, one explains how-possibly to dispel a questioner's puzzlement at having witnessed or been told of some event. One can do away with this psychological element, however, by explicating surprise in epistemic terms; i.e., as a prima facie tension between the fact to be explained and the questioner's body of knowledge absent some additional piece of information. This latter is what is provided by a how-possibly explanation. Other approaches to modifying Dray's view so that it is more conformant to ontic conceptions of explanation have also been considered (see Persson, 2012).

incomplete how-actually explanation.[5] Explaining how-possibly, for Forber, is not this but a kind of formal inquiry: given a set of relevant background assumptions, one deduces (e.g., via computer simulation) a particular set of outcomes reachable from them. For instance, let the assumptions consist of known biological laws relevant to a particular population, plus a specification of a set of variable parameters, and let the different outcomes represent various genotypes associated with that population. Then, when one runs such a simulation, the different paths by which it arrives at a particular outcome carve up the possibility space for that outcome—they represent a set of how-possibly explanations corresponding to it.[6] Explaining how-actually, in contrast, is a form of empirical inquiry: its aim is to determine which of these possible explanations is the actual one.

A further mode of how-possibly explanation, described by Persson, "aims to establish the existence of a mechanism by which X could be, and was, generated without filling in all the details" (Persson, 2012, 275). Key to Persson's conception is the empirical determination of some actually existing mechanism responsible for X. It is thus distinct from Resnik's conception of how-possibly explanation as inadequately supported how-actually explanation. It is also distinct from Forber's conception of how-possibly explanation, which recall is not a form of empirical inquiry at all. Nevertheless it is how-possibly and not how-actually explanation because, although one describes an actual mechanism responsible for X,

[5]One could be forgiven for thinking this merely a dispute over labels. In Forber's defence, his view is more in the spirit of Dray's original account, who recall, viewed how-possibly and how-actually questions as logically different.

[6]Forber distinguishes between *global* and *local* how-possibly explanations. The former utilise highly idealised background assumptions. The latter are directed at real populations and utilise richer, empirically grounded, assumptions.

information is missing from our description of the mechanism which would allow us to determine the precise (typically causal) pathway by which X was brought about. Thus one has not given an account of how the event *actually* occurred.

In the modes of explaining how-possibly identified so far, the questioner is required, or at any rate it is perfectly sensible for her, to continue on to ask the how-actually question. Thus for Hempel she (at least in interesting cases; see Hempel 1965, 429) will not be fully satisfied until she answers the how-actually question. On Resnik's conception, how-possibly questions are just how-actually explanations that have not been adequately confirmed, and it goes without saying that we should try and confirm them. For Persson it is not confirming but "filling in" of the mechanism which remains to be done. On Forber's conception, explaining how-possibly is in no way inferior to explaining how-actually, yet both play an essential role in our inquiries into phenomena. For Dray, pace Hempel, sometimes one is thoroughly satisfied with a how-possibly explanation. Nevertheless it is perfectly sensible to go on and ask exactly how the centre fielder caught the ball once he was at the top of the ladder.

The kind of how-possibly explanation I describe in the next section bears certain resemblances to both Persson's and Forber's conceptions. As in Persson's conception it involves the description of some mechanism actually responsible for producing an outcome. Unlike in Persson's conception, in this case there are no relevant details of the mechanism left to fill in. On the other hand, the way the mechanism explains, as in Forber's conception, is that it carves out a particular possibility space for an outcome. But unlike all of the conceptions reviewed above, once one has answered the how-possibly question in this case, it is doubtful that a how-actually explanation can give us anything more of substance.[7]

[7]Which of these conceptions is right? With Persson I would maintain these are all legitimate senses of explaining how-possibly. Which one is 'correct' will be relative to the

3 Explaining Quantum Speedup

A basic distinction, in Computational Complexity Theory, is between those computational problems amenable to an efficient solution in terms of time and/or space resources, and those that are not. Easy (or 'tractable', 'feasible', 'efficiently solvable', etc.) problems have solutions which involve resources bounded by a polynomial in the input size, n (n is typically the number of bits used to represent the input). Hard problems are those which are not easy; i.e., they require resources that are 'exponential' in n, i.e., that grow faster than any polynomial in n (Nielsen and Chuang, 2000, p. 139).[8] For example, a problem which requires $\approx n^c$ time steps to solve in the worst case (for some constant c) is polynomial in n and thus tractable according to this definition. A problem that requires $\approx c^n$ steps, on the other hand, is exponential in n therefore intractable according to this definition.

'Quantum speedup' refers to the fact that some computational problems can be solved exponentially faster with a quantum computer than with any known classical computer.[9] For example, the fastest known classical algorithm for factoring the product of two unknown primes is exponential in n. Shor's quantum algorithm, astoundingly, solves the problem in polynomial time (Nielsen and Chuang, 2000, 216). But while the fact of quantum speedup is almost beyond doubt,[10] its source is still a matter of debate within the scientific community.

specific question asked.

[8]The term 'exponential' is being used rather loosely here. Functions such as $n^{\log n}$ are called 'exponential' since they grow faster than any polynomial function, but they do not grow as fast as a true exponential such as 2^n.

[9]Research into quantum computing is still largely in the theoretical stage. However there is good reason to be optimistic that practical implementations will be realised eventually (see Aaronson, 2013, Ch. 15).

[10]There is currently no proof that the class of problems efficiently solvable by quantum

```
void SelectionSort(int intsToSort[], int lengthOfList) {
  // Declare list indices:
  int i, j, indexOfLowestNum;
  // For each position in the list,
  for (i = 0; i < lengthOfList − 1; i++) {
    // provisionally assert that it points to the lowest number,
    indexOfLowestNum = i;
    // and then for each of the other list positions,
    for (j = lengthOfList − 1; j > i; j−−) {
      // if the number pointed to by it is less than the number
      // pointed to by indexOfLowestNum,
      if (intsToSort[j] < intsToSort[indexOfLowestNum]) {
        // then make this the new provisional minimum index.
        indexOfLowestNum = j;
      }
    }
    // At the end of the ith iteration, put the number that is in the
    // indexOfLowestNum position into the ith position (and vice versa).
    Swap(&intsToSort[i], &intsToSort[indexOfLowestNum]);
  }
}
```

Figure 1: A set of instructions (in C) implementing the 'SelectionSort' solution to the problem of sorting a list of given integers.

According to Fortnow (2003), for instance, the explanation of quantum speedup lies in the ability of quantum systems to exhibit 'interference'. For Ekert and Jozsa (1998), on the other hand, it is their ability to exhibit 'entanglement'.

We will consider these explanations in a little more detail later. For now let us ask: what kind of question is one asking when one asks to have quantum speedup explained? It is clearly an ontic question, for it aims to identify particular characteristics of physical systems. It is also a how question, if anything is, for it asks, specifically, for the distinctive mechanism by which quantum computers operate. It remains to consider whether it is a how-possibly or a how-actually question.

Consider figure 1. Depicted there is an instance of the 'SelectionSort' algorithm for sorting

computer is larger than the class efficiently solvable by classical computer, however other results make the truth of this statement very likely (see Aaronson, 2013, Ch. 10).

a given list of integers. If given, say, the numbers (25, 12, 13, 19, 8), then after a certain time a computer running this code will produce: (8, 12, 13, 19, 25). Now if we examine the algorithm, we notice that in the worst case (indeed, in any case) there will be $n - 1$ comparisons in the first iteration, $n - 2$ comparisons in the second, $n - 3$ in the third, and so on (where n is the number of list items). This gives a total of $n(n - 1)/2$ comparisons; thus our total worst-case running time is proportional to n^2. SelectionSort is not the only algorithm for sorting integers. Both faster and slower algorithms exist. For instance, the 'MergeSort' algorithm has a worst-case running time $\propto n \log(n)$ (Mehlhorn and Sanders, 2008).

Suppose we are comparing the running times of various algorithms. I feed some random permutation of a list of n integers to my algorithm (which implements SelectionSort) and you feed one to yours (which implements MergeSort). After a time we both obtain the list in sorted order. Yours finishes after $k \leq n \log(n)$ steps. Mine finishes after $n \log(n) < l \leq n^2$ steps. What is the explanation for this difference in performance? Well, one way to explain it is just to point to the differences in the code for the two algorithms. But what does this code represent? Certainly it does not represent some one particular linear causal sequence of transitions, for the **if** as well as the **for** loops encode conditional statements. Rather, it represents a *space of possibilities*: a set of pathways by which the computer can arrive at a particular result. It turns out that the pathways available to a computer implementing MergeSort allow a solution to be reached in fewer time steps than the pathways available to one implementing SelectionSort.

Something similar can be said when comparing different classes of machine. Consider, for instance, figure 2. 'State transition diagrams' such as these are essentially just another way of representing algorithms, although the representation they afford is somewhat 'closer to the hardware', so to speak. The machine depicted is an example of a deterministic finite automaton (DFA). It is not the only kind of computing machine. There are also, for instance,

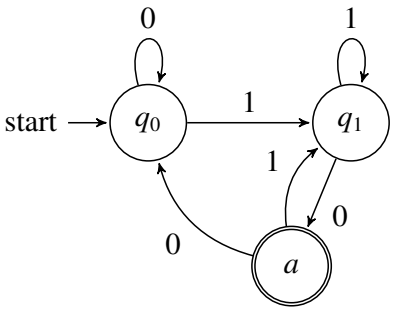


Figure 2: A state diagram representation of a deterministic finite automaton (DFA). Binary strings of variable length are input to the automaton. They are 'accepted' if the machine is found to be in the state *a* after the last character has been read. This particular machine will accept any string ending in '10'.

nondeterministic finite automata, deterministic and nondeterministic 'pushdown' automata, and deterministic and nondeterministic Turing machines, to name a few. To define these various classes of machine, we describe the possible states and state transitions which they are capable of. For example, DFAs are characterised by a finite set of states, deterministic transitions between states, and the lack of any form of external storage (see Martin, 1997).

Given our characterisations of different types of machine, we can inquire about the set of problems computable by the machines of a particular class. It turns out, for example, that DFAs are severely limited with respect to the class of problems they are capable of solving, while Turing machines, in contrast, are capable of solving any effectively calculable function. We can similarly ask about the resources required to solve particular classes of computational problems by machines of a particular sort. We can ask, for instance, about the class of problems solvable by a deterministic Turing machine in polynomial time, about those solvable by a nondeterministic Turing machine in exponential time, and so on. Answering these and other similar questions will involve appealing to the states and to the state

transitions which are possible for a particular class of machine. This state space, we will say, allows us to construct such a machine to realise an algorithm that will solve the problem in a given amount of time.

The question, "what is the source of quantum speedup?", is a question of just this sort. Quantum computers are just another type of computational machine, and just as for Turing machines and DFAs, quantum computers have associated with them a particular space of states and a particular way of transitioning between states. In order to answer the question "what is the source of quantum speedup?", therefore, we will appeal precisely to the quantum mechanical state space and to the allowable transitions within it, and we will consider these in comparison to the space of states and state transitions possible for a classical computer. And when we do so we will be explaining how-possibly.

We see this, in fact, when we examine the approaches that those in the scientific community have taken to answering this question. Consider Fortnow (2003), for example, who develops an abstract mathematical framework for representing both the computational complexity classes associated with classical and with quantum computing. In Fortnow's framework, both kinds of computation are represented by transition matrices which determine the allowable transitions between possible configurations of a particular kind of machine. To represent the quantum case, Fortnow allows matrix entries to be negative as well as positive, while for the classical case they may only be positive. As a result, in the quantum case matrix entries will sometimes cancel each other out when summed; not so in the classical case. Fortnow shows that this suffices to capture the computational complexity classes associated with classical and quantum computing. According to Fortnow, this means that the fundamental difference between quantum and classical computation is that in quantum computation there can sometimes be interference between computational paths: "The strength of quantum

computing lies in the ability to have bad computation paths eliminate each other thus causing some good paths to occur with larger probability" (Fortnow, 2003, p. 606).[11] Ekert and Jozsa (1998), on the other hand, argue that the fact that quantum systems can sometimes be in entangled states yields a state space for combined quantum systems that is exponentially larger than the state space associated with combined classical systems.[12] And while it is possible to, in a roundabout way, simulate this larger state space with a classical computer, the resource cost of doing so scales exponentially (ibid., 1771).[13]

[11]In the 'many worlds' interpretation (Hewitt-Horsman, 2009), of course, all of these paths would be actual and not merely possible. Perhaps. But I do not think it prudent to hinge one's views on scientific explanation on a particular interpretation of quantum mechanics (further, see Cuffaro 2012 for some strong reasons to be skeptical of the many worlds view in the context of quantum computing). This is moot in any case. When a quantum computer finds itself in a state like $|\psi\rangle = |000\rangle_i|f(000)\rangle_o + |001\rangle_i|f(001)\rangle_o + \cdots + |111\rangle_i|f(111)\rangle_o$ we do not, in order to make sense of Fortnow's analysis, need to take the terms in this superposition to represent either actual or possible computational paths. Rather, what is important is only that the possible states of a quantum computer, *unlike* those of a classical computer, include superpositions like $|\psi\rangle$ which have interfering terms.

[12]A system of two or more particles is said to be entangled when one cannot describe one of the particles in the system without referring to all of the others.

[13]There is disagreement here between Fortnow and Ekert and Jozsa. Does this undermine my view? I would say not. We have here two potential how-possibly explanations of quantum speedup. Further empirical research, presumably, will help to decide which of these how-possibly explanations is correct. One might investigate, for instance, whether cases exist in which a quantum algorithm is efficiently classically simulable despite the fact that it utilises entangled states.

But is this really explaining how-possibly? Isn't it the case, one might object, that an algorithm like SelectionSort just is a description of how a system actually goes about solving a problem, and likewise that a description of the state space and state transitions associated with a particular class of system just is a description of the actual resources used by those systems? These mechanisms are *actually* being employed, are they not? Of course this is true. In the same way, Dray's centre fielder *actually* used the ladder to make the catch that he did. But that is not a how-actually explanation, and neither are these. For Dray the role played by a description of the ladder in the explanation of the fielder's catch is to dispel the questioner's puzzlement regarding how the catch was possible, without explaining exactly how it happened (or why it was necessary). But why does the ladder dispel her puzzlement? It does so because pointing out that a ladder was present opens up a whole new range of possibilities for the questioner that simply weren't there before. Likewise in the cases we are considering: the algorithms, or the state spaces, as the case may be, explain by explicitly specifying the set of possibilities open to them.

But is the how-possibly explanation fully satisfactory? Shouldn't we feel the urge to continue our investigation until we have found the actual path taken by the computer through its state space? Let us consider what this would mean in the case of the performance comparison between SelectionSort and MergeSort. In this case we would presumably (assuming the precise input value was known) produce a how-actually explanation by giving a detailed description of the state of the computer after each time step, and in this way we would see exactly how it was that mine took l steps and yours took k. And yet, without referencing the possibility spaces carved out by each algorithm—the alternatives encoded in the conditional statements—it is hard to see how such an answer can be very informative with respect to the actual question that was asked. At best it is to answer a very different question

than the one originally asked. But in the context of a discussion of the performance characteristics of different types of computer it is not clear what an answer to such a question will add to our understanding of these processes. Such an answer, in that context, seems to do little more than restate the original question. The information about the performance characteristics of my and your computer is most crucially contained in the description of their possibility spaces. Such questions are therefore most appropriately answered by appealing to those possibility spaces; i.e., they are most appropriately answered with how-possibly explanations.

4 Conclusion

The kind of how-possibly explanation I have described in this paper bears certain resemblances to both Forber's and to Persson's conceptions of explaining how-possibly. As in Persson's conception, an explanation of the comparative performance characteristics of quantum and classical computers involves, I have argued, a description of the actual mechanisms associated with these machines. The description of a mechanism serves to carve out a particular possibility space for a machine, and as on Forber's view, this possibility space plays a crucial role in a how-possibly explanation of the computationally relevant characteristics of particular observed outcomes. I have further argued that, unlike other interesting examples previously given in the literature on this form of explanation, once one has answered this how-possibly question it is doubtful that continuing on to ask a how-actually question can yield anything more of substance.

The kind of how-possibly question I have described here is characteristic, not only of the inquiry into the source of quantum speedup, but of the science of computability and computational complexity generally—and not just this. Algorithmic processes abound in

nature: in biological systems, in cognitive systems, and also in physical systems, to name but a few. Questions regarding their comparative performance characteristics likewise abound. And I would argue, if I had the space, that all of these questions are most appropriately answered by explaining how-possibly.

References

Aaronson, Scott. *Quantum Computing Since Democritus*. New York: Cambridge University Press, 2013.

Cuffaro, Michael E. "Many Worlds, the Cluster-state Quantum Computer, and the Problem of the Preferred Basis." *Studies in History and Philosophy of Modern Physics* 43 (2012): 35–42.

Dray, William. *Laws and Explanation in History*. Oxford: Oxford University Press, 1957.

Ekert, Artur, and Richard Jozsa. "Quantum Algorithms: Entanglement-enhanced Information Processing." *Philosophical Transactions of the Royal Society A* 356 (1998): 1769–1782.

Forber, Patrick. "Confirmation and Explaining How Possible." *Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (2010): 32–40.

Fortnow, Lance. "One Complexity Theorist's View of Quantum Computing." *Theoretical Computer Science* 292 (2003): 597–610.

Hempel, Carl G. *Aspects of Scientific Explanation And Other Essays in the Philosophy of Science*. New York: The Free Press, 1965.

Hempel, Carl G., and Paul Oppenheim. "Studies in the Logic of Explanation." *Philosophy of Science* 15 (1948): 135–175.

Hewitt-Horsman, Clare. "An Introduction to Many Worlds in Quantum Computation." *Foundations of Physics* 39 (2009): 869–902.

Martin, John C. *Introduction to Languages and the Theory of Computation*. New York: McGraw-Hill, 1997, second edition.

Mehlhorn, Kurt, and Peter Sanders. *Algorithms and Data Structures*. Berlin: Springer, 2008.

Nielsen, Michael A., and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press, 2000.

Persson, Johannes. "Three Conceptions of Explaining How Possibly—and One Reductive Account." In *EPSA Philosophy of Science: Amsterdam 2009*, edited by Henk W. de Regt, Stephan Hartmann, and Samir Okasha. Dordrecht: Springer, 2012, 275–286.

Reiner, Richard. "Necessary Conditions and Explaining How-Possibly." *The Philosophical Quarterly* 43 (1993): 58–69.

Resnik, David B. "How-Possibly Explanations in Biology." *Acta Biotheoretica* 39 (1991): 141–149.

Salmon, Wesley C. "Scientific Explanation: Three Basic Conceptions." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1984 (1984): 293–305.

———. "Four Decades of Scientific Explanation." In *Scientific Explanation (Minnesota Studies in the Philosophy of Science, Volume XIII)*, edited by Philip Kitcher, and Wesley C. Salmon. Minneapolis: University of Minnesota Press, 1989, 3–219.

Woodward, James. *Making Things Happen*. Oxford: Oxford University Press, 2003.

The Mathematics of Causal Capacities

David Danks

Departments of Philosophy & Psychology

Carnegie Mellon University

ddanks@cmu.edu

Abstract

Models based on causal capacities, or independent causal influences/mechanisms, are widespread in the sciences. This paper develops a natural mathematical framework for representing such capacities by extending and generalizing previous results in cognitive psychology and machine learning, based on observations and arguments from prior philosophical debates. In addition to its substantial generality, the resulting framework provides a theoretical unification of the widely-used noisy-OR/AND and linear models, thereby showing how they are complementary rather than competing. This unification helps to explain many of the shared cognitive and mathematical properties of those models.

1. Introduction

In many scientific domains, one finds models focused on causal influences that function (at least somewhat) independently of one another. For example, cognitive models are typically expressed in terms of distinct cognitive processes that have no direct influence on one another's functioning, and so can proceed independently, whether sequentially or in parallel. As just one instance, many theories of categorization posit that people first perceive the relevant stimulus, then judge its similarity to various known categories, and finally use those similarity judgments to generate a behavioral response. These processes obviously matter for one another; the output of the perceptual process, for example, is the input to the similarity judgment process. But in essentially all similarity-based cognitive theories of categorization, the *functioning* of one process is assumed to be largely independent of the *functioning* of the other processes. The "inner workings" of the perceptual process are assumed to be irrelevant to the way that similarity judgments are made; the only influence of the former on the latter is the particular information that it outputs.

More generally, scientific models and theories frequently divide the world into distinct processes (typically, causal ones) such that the operation of one process has minimal dependence on—in the best case, true independence from—the operations or states of other processes. Probably the clearest articulation of this picture is based on the notion of causal capacities (Cartwright 1989, 1999, 2007; Martin 2008; see also Heil 2005), but similar ideas can be found in many writings on mechanisms (in the spirit of Machamer, Darden, & Craver, 2000). In this paper, I focus on such independent causal influences; for convenience, I will refer to them as 'capacities', but this term should be understood broadly. The basic idea is that capacities are just those causal powers that a cause *C* has purely by virtue of being a *C*; causal capacities are

"something they [the causes] can be expected to carry with them from situation to situation" (Cartwright 1989, 145). That is, capacities inhere in *C* rather than arising from the particular situation, and so their operation should be relatively unaffected by other processes in the system. This (almost) independence is exactly what enables the construction of "nomological machines" (Cartwright 1999, 2007) that generate the regularities—some contingent, some law-like—that we observe and manipulate.

The philosophical literature on causal capacities and mechanisms has largely focused on questions that are metaphysical (e.g., are they basic/fundamental features of the world?) or epistemological (e.g., can we discover capacities from observational or experimental data?). I here consider a representational question: is there a natural, privileged representational framework for systems in which the causal influences[1] are independent[2] of one another (i.e., each does not depend on the values, operations, or status of the others)? There is enormous variety in the world, and so any representational framework inevitably simplifies or is sometimes not applicable. My interest here is in a representational framework that applies to the "standard" or "ordinary" cases, and so can function as a default framework; I use the terms 'natural' and 'privileged' to refer to such a framework. One might think that there obviously can be *no* such

[1] For simplicity, I assume that each independent influence corresponds to a single cause, as multiple (interactive) causes can be merged into a single, multidimensional, factor.

[2] This independence should not be confused with (a) statistical independencies that can be used to (sometimes) infer causal structures from data (Spirtes, Glymour, and Scheines 2000); or (b) the idea of 'modularity' to refer to causal connections that can be separately intervened upon (Hausman and Woodward 1999, 2004; Cartwright 2002).

privileged representation, as the independence property seems too weak for this task, but that response turns out to be mistaken.

2. The Mathematics of (a Special Case of) Causal Capacities

2.1. The Noisy-OR/AND Model

Assume that we have a set of (possible) causes C_1, …, C_n and a target effect E. The functioning of C_i's capacity is supposed to inhere in C_i, and so the causal strength or influence of C_i should be representable without reference to the states of the other variables. In particular, C_i's impact on E should not depend on the state or causal strength of C_j, and it should be monotonic in C_i; in particular, even if the quantitative impact is not constant across values of E (due to, e.g., saturation of E), the valence should not depend on E's value. Finally, for mathematical tractability, I assume that each variable's possible values can be represented as numbers, though each variable can have its own scale; this is a trivial assumption when the variables are binary (i.e., two-valued), but is non-trivial in other cases (e.g., there is no privileged way to map *red*, *green*, and *blue* to numbers).

Consider the special case situation in which all factors—causes and the effect—can be represented as binary variables. For this case, a privileged mathematical framework (with origins in 19th century mathematics) has been developed in machine learning and cognitive psychology (Good 1961; Srinivas 1993; Heckerman and Breese 1994, 1996; Cheng 1997; Glymour 1998; Cozman 2004). Suppose that we have a single generative (binary) cause C_1 of the (binary) effect E, and so E occurs when (and only when) C_1 is present and the capacity of C_1 is active, where w_1 is the strength of that capacity. Thus, we immediately derive $P(E) = w_1 \times \delta(C_1)$, where $\delta(X) = 1$ if X is present, 0 if X is absent. If we have a second generative cause C_2 of E, then E occurs when

(and only when) either C_1 or C_2 generates it, where the 'or' is non-exclusive. Thus, we have $P(E) = w_1\delta(C_1) + w_1\delta(C_2) - w_1\delta(C_1)w_2\delta(C_2)$; that is, the probability of E is just the sum of the probabilities that it is caused by one cause, minus the probability that both caused it (in order to account for that case being "double-counted" in the sum of the first two terms). More generally, if we have n distinct, independent generative causes, then the resulting expression for $P(E)$ is the "noisy-OR" model (Good 1961; Kim and Pearl 1983; Pearl 1988; Srinivas 1993; Heckerman and Breese 1994; Cheng 1997; Glymour 1998):

$$P(E|C_1, \ldots, C_n) = 1 - \prod_{i=1}^{n}(1 - w_i\delta(C_i)) \tag{1}$$

In a noisy-OR model, E is an OR-function of the different causes, but with cause-specific "noise" (understood instrumentally) that probabilistically makes that cause's capacity inactive. Thus, the probability that E occurs is just the probability that at least one present cause has an active capacity. Moreover, equation (1) is uniquely privileged: it is the *only* equation for purely generative binary causes with distinct causal capacities (i.e., independent causal influences) that satisfies various natural properties (Cozman 2004).

Of course, not all causes are generative; we are often interested in causes that *prevent* the effect from occurring. If a preventive cause P interferes with the functioning of only one specific generative cause G,[3] then P has the (mathematical) impact of reducing G's causal strength and so we can combine their causal capacities. We cannot do the same for preventers that apply to all generators equally; such preventers operate as (noisy, probabilistic) "switches" that control whether any generative cause can be active at all. That is, E occurs when (and only when) at least

[3] An ambiguity lurks here between "prevention as blocking" and "prevention as reducing," but I postpone discussion of this ambiguity until later in this section.

one generative cause's capacity is active and none of the preventive causes' capacities is active. This relationship is captured by a "noisy-OR/AND" model, since the generative causes combine in a noisy-OR function, whose result is then combined with a noisy-AND function for the preventive causes (i.e., the effect occurs only if a generator is active AND P_1 is not active AND … P_m is not active):

$$P(E|C_1,\ldots,C_n,P_1,\ldots,P_m) = \prod_{j=1}^{m}(1-w_i\delta(P_j))\Big[1-\prod_{i=1}^{n}(1-w_i\delta(C_i))\Big] \quad (2)$$

This equation provides (arguably) the most natural representation of causal capacities, both generative and preventive, that exert independent causal influence (Srinivas 1993; Heckerman and Breese 1994, 1996; Lucas 2005). Moreover, there is substantial empirical evidence that humans preferentially represent causal systems as functioning according to equation (2) (Cheng 1997; Holyoak and Cheng 2011; Danks 2014).[4]

2.2. Resolving Ambiguities

Although there is great value in this mathematical framework, the restriction to binary variables is significant, as there are many cases in which the influence of a causal capacity depends in part on the factor's magnitude or intensity, or the effect can exhibit fine degrees of meaningful variation. Before generalizing the noisy-OR/AND model to many-valued variables, however, we must clarify two key conceptual (though not mathematical) ambiguities.

[4] The connection between psychological theory and capacities is unsurprising, as Cheng's (1997) causal power theory in cognitive psychology was explicitly modeled on Cartwright's (1989) capacity account of causation.

Mathematically speaking, binary variables are simply those with two possible values. When talking about causal capacities, however, a more specific interpretation is typically intended: factors can be "present" vs. "absent" or "on" vs. "off"; capacities can be "active" vs. "inactive". These interpretations provide a natural value ordering, as shown by the standard practice of mapping "present" to the value of 1 and "absent" to the value of 0.[5] More generally, we typically understand the "absent" or 0 value to be the *lower bound* of the possible values for that variable. At the same time, the zero value in the context of causal capacities almost always serves as the *baseline* value: it is the value that E would have if nothing influenced it. This second role of the zero value is clear in the mathematics of the noisy-OR/AND model, as $P(E = 0 \mid$ all generative causes are absent$) = 1$. That is, the standard model of (binary) causal capacities assumes that absence is the appropriate "uncaused" state for E.[6]

These two different roles for zero—lower bound and baseline value—are conceptually distinct and empirically distinguishable. For example, in most terrestrial environments, the baseline value for *Oxygen in Room* (i.e., the value it has when represented causes are all inactive) is "present," not "absent." We can represent this different baseline value in the noisy-OR/AND model, but only through a mathematical trick (namely, a very strong, always-present generative cause). A better solution would be to allow the lower bound and baseline to diverge. This

[5] This particular mapping could obviously be reversed without any change in substantive content, though 'lower bound' and 'upper bound' would need to be swapped in what follows.

[6] One might worry that "uncaused states" are impossible. However, if causes function independently, then it is at least theoretically possible for none to be active at a moment in time. More generally, *any* model with independent causal influences yields a baseline value, even if it is only ever theoretically (rather than empirically) realized.

generalization does not matter for cases with only binary variables, as any model with variables whose baseline is 1 can be translated into a model in which all baselines are 0. Outside of this special case, however, the baseline value plays a distinct mathematical role, and so any model of causal capacities that allows for more-than-binary variables (such as the one developed in Section 3) must distinguish conceptually between the lower bound value and the baseline value.

The multiple roles played by zero point towards the other important ambiguity in the standard noisy-OR/AND model of causal capacities. Because the zero value is both the lower bound and the baseline, there are two different ways to prevent, or make *E* less likely. First, the preventer could stop generative causes from exerting their usual influence. These *blockers* serve to keep the effect variable closer to its baseline value, as they (potentially) eliminate causal influences that drive the effect away from baseline. Preventive causes in the noisy-OR/AND model are usually understood in this way. A second way of "preventing" is to move *E* towards its lower bound. These *reducers* are the natural opposite of standard generative causes, as they shift *E* downwards while generators shift *E* upwards. The important distinction here is whether the preventer influences the effect directly (i.e., is a reducer), or indirectly through the elimination of other causal influences (i.e., is a blocker).

As a practical example, suppose *Heart Rate* is our effect variable. There are many generative causes that increase heart rate, such as stress or exercise. Beta blockers and other anxiety-reducing medications function as blockers, as they prevent (some of) those generative causes from having any influence while not suppressing *Heart Rate* below its natural baseline (for that individual). In contrast, most anesthetics are reducers of *Heart Rate*, as they actively slow the heart, potentially even below its natural baseline, depending on exactly which causes are active. Of course, if we model *Heart Rate* as simply "low" or "high" (where "low" is the baseline), then

these two different types of drugs will appear indistinguishable. The importance of distinguishing reducers from blockers becomes apparent only when we move to situations in which the lower bound and baseline values need not coincide.

Before turning to the fully general mathematical framework for causal capacities, we must address a potential ambiguity about a capacity's "causal strength" w_i. The standard interpretation in the noisy-OR/AND model is that w_i expresses the probability that the capacity is "active," where an active cause deterministically produces the effect (unless a suitable blocker is also active). This interpretation is inappropriate when causes are more than binary, as "probability of activation" neglects the (presumed) importance of the magnitude of the cause variable.[7] Instead, we will understand w_i (for generators and reducers) for a capacity C_i to be the expected change in E's value when C_i increases by one unit and every other factor is at its baseline value. That is, w_i is computed by starting in the state in which every causal factor is at baseline, and then determining the *expected* change in E when C increases by one unit.[8] This interpretation implies that w_i depends on C_i's scale, but this should be expected given the predictive function of causal strengths. Notice that, if all causes and the effect are binary, then the expected change and probability of activation interpretations of w_i are mathematically identical. The expected change

[7] We can retain the "probability of activation" interpretation if the effect is the only many-valued variable, in which case the natural representations are noisy-ADD or noisy-MAX functions (Heckerman and Breese 1996).

[8] If causal strength depends on C's value, then the choice to measure from C's baseline is potentially a substantive one. However, since we assume causes have independent monotonic influences, we can always transform the scale for C so that E is a linear function of C's value.

interpretation, however, also naturally applies to systems in which some factors can take on more-than-two values, and so I use it in the next section.

3. A General, Privileged Mathematical Representation

Now that we have done the necessary conceptual clarification, we can develop a general, privileged mathematical representation of causal capacities when the causes and effect need not be binary. Throughout, I use lower-case letters to denote the value of a variable; for example, e is the value of the effect E. Without loss of generality, we can assume E's baseline value is zero and $e \in [-L, U]$, where at least one of L, U is greater than zero (else E is always zero). Note that the baseline can be the same as the lower bound ($L = 0$, $U > 0$); same as the upper bound ($L > 0$; $U = 0$); or a strictly intermediate value (L, $U > 0$). As noted above, three different types of causal capacities must be incorporated into the mathematical framework: generators G_i and reducers R_j that (probabilistically) increase and decrease the value of E, respectively; and blockers B_k that (probabilistically) prevent any other causal capacities from influencing E. For all three types of causes, their values must also be able to range over more than just $\{0, 1\}$. For mathematical convenience, we represent the "inactive" state of each cause by 0, so that the influence on E (when only C is active) is the product of C's magnitude (i.e., its distance from zero) and its causal strength (i.e., the expected change in E given that the cause increased by one unit).[9]

Consider first the case with only generators G_i with values g_i. In this situation, E can only be pushed upwards from the baseline, and so $e \in [0, U]$. The natural mathematical framework

[9] Recall from fn. 8 that the independent causal influences have all been transformed so that they are linear (in C) influences of E, and so this product for a single generator or reducer is always less than the relevant upper or lower bound, respectively.

simply uses normalization to convert this case to (a continuous version of) the noisy-OR model: (i) "normalize" *E* and the causal strengths to the [0, 1] interval; (ii) use the uniquely privileged (Cozman 2004) noisy-OR model; and then (iii) transform the result back to the [0, *U*] interval. The noisy-OR/AND model was defined in equations (1) and (2) in terms of the probability of *E* given its causes, but we can (and should, in the present context) instead regard those equations as providing the expectation of *E*. The natural, privileged mathematical representation for the expectation of *E* in this situation is thus:[10]

$$\mathbb{E}(E) = U[1 - \prod_{i=1}^{g} \frac{U - w_i g_i}{U}]$$

Since reducers are naturally understood as "negative generators," we can model the impact of a set of reducers R_j with values r_j in the same way, though their "normalization" is relative to *L* rather than *U*. The resulting expectation of *E* is simply the difference between these (normalized and combined) influences:

$$\mathbb{E}(E) = U[1 - \prod_{i=1}^{g} \frac{U - w_i g_i}{U}] - L[1 - \prod_{j=1}^{r} \frac{L - w_j r_j}{L}]$$

Finally, blockers B_k with values b_k fill the role of preventers in the noisy-OR/AND model of equation (2): the (probabilistic) activation of their causal capacities prevents the expression of any other causal capacities, and so they act as a probabilistic "switch" on the previous equation. The causal strengths of the blocking capacities are thus best understood as "increase (per unit change in the blocker) in probability of complete blocking." The resulting full mathematical equation is:

[10] I show below that this equation is well-behaved even when $U = +\infty$.

$$\mathbb{E}(E) = \prod_{k=1}^{b}(1 - w_k b_k)\Big[U[1 - \prod_{i=1}^{g}\frac{U - w_i g_i}{U}] - L[1 - \prod_{j=1}^{r}\frac{L - w_j r_j}{L}]\Big] \quad (3)$$

Equation (3) is the natural generalization of the noisy-OR/AND model to cases with many-valued variables and distinct baseline and lower bound for *E*. It thus provides the privileged mathematics of causal capacities for precisely the same reasons as the noisy-OR/AND model for the special case of binary variables. To see that it provides such a generalization, consider the special case that was the focus of the previous section: $L = 0$, $U = 1$, and all of the causal factors are restricted to $\{0, 1\}$. Since *L* is equal to the baseline, there are no "reducing" causal capacities: for any putative reducer *R*, the expected change in *E* from a unit change in *R* (when all other causes are absent) is always zero, and so w_R is always zero. And since the causal factors are restricted to $\{0, 1\}$, the b_k and g_i variable values can be replaced with delta functions. The resulting equation (when we substitute in *U* and *L*) is simply equation (2), the noisy-OR/AND model. That is, the equations and claims of the previous section are all special cases of the generalization provided here.

Equation (3) provides a privileged mathematics for arbitrary variable ranges and causal capacities, in the sense (previously articulated) that it captures the plausible, intuitive features of "standard" cases, and therefore can serve as a natural default representational framework. It is particularly interesting to consider another special case. Suppose $E \in [-\infty, +\infty]$ and (for the moment) that there are no blockers. It is not obvious how to use equation (3) in this situation, since direct substitution of *L* and *U* yields infinities throughout the equation. If we instead

consider the limit of equation (3) as L and U go to infinity, we find that the expectation of E is given by:[11]

$$\mathbb{E}(E) = \sum_{i=1}^{g} w_i g_i - \sum_{j=1}^{r} w_j r_j \tag{4}$$

That is, the natural mathematical equation for (the expectation of) E in this special case is simply a linear function of the causal capacities. Having seen equation (4), it is straightforward to incorporate blockers, as that initial product term will simply act to globally attenuate the linear impact (on the expectation of E) of the generators and reducers.

Equation (3) provides a measure of unification to equations (2) and (4): despite their substantial mathematical differences, both noisy-OR/AND and linear models are special cases of the more general, privileged mathematical characterization of causal capacities. That is, this framework suggests that noisy-OR/AND and linear models have the same conceptual and mathematical basis, and the different models arise simply based on whether the variables are binary or continuous/real-valued. In particular, this unification helps to explain why so many mathematical results that hold for linear models also hold for noisy-OR/AND models, and vice versa. For example, the conditions for model parameter identifiability are essentially the same for noisy-OR/AND models (Hyttinen, Eberhardt, and Hoyer 2011) and linear models (Hyttinen, Eberhardt, and Hoyer 2012). Similarly, we find basically the same conditions and statistical tests

[11] *Proof sketch*: For the generators in equation (3), separate the fraction terms into differences and expand the product to yield: $U\,[1 - (1 - \Sigma\,(w_i g_i / U) + \mathbf{C})] = [\,\Sigma\, w_i g_i - U\mathbf{C}]$, where $\mathbf{C}$ is the rest of the product expansion. Every term in $\mathbf{C}$ has at least U^2 in the denominator, and so as $U \to +\infty$, $U\mathbf{C} \to 0$. Thus, as $U \to +\infty$, we are left with only the sum. The same reasoning yields the sum for reducers.

for discovering an unobserved common cause of multiple observed effects given either a noisy-OR/AND model (Danks and Glymour 2001; Pearl 1988) or a linear model (Spirtes et al. 2000). This overlap in the models' mathematical properties is much less surprising given that they (arguably) derive from a single, more general equation (though their properties are not identical, since the different variable value ranges do sometimes matter).

This mathematical connection can also provide us with insights into human cognition. I earlier noted that the noisy-OR/AND model emerged partly from work in cognitive psychology on one "natural" way that people seem to represent causal strengths in the world, at least when we have binary causes and effects (Cheng 1997; Danks 2014; Holyoak and Cheng 2011). At the same time, there are competing theories of human causal learning—variants of the Rescorla-Wagner model and its long-run counterpart, the conditional ΔP theory (Danks 2003)—in which people represent causal capacities as combining linearly (Danks 2007). Relatedly, there is a long history of psychological research on function approximation that has shown that people find linear functions easier to learn (e.g., McDaniel and Busemeyer 2005; DeLosh, Busemeyer, and McDaniel 1997; and references therein), and even have a significant bias in favor of understanding the world in terms of linear functions (Kalish, Griffiths, and Lewandowsky 2007). Equation (3) provides a measure of theoretical unification for these disparate psychological results: noisy-OR/AND and linearity are not theoretical competitors, but rather different aspects of the same general assumptions or preferences about causal capacities. That is, we need not ask whether noisy-OR/AND *or* linearity is correct, since each is the natural representation *for a particular domain of variable values*.[12]

[12] This observation suggests that people in causal learning experiments might systematically shift between noisy-OR/AND and linearity based solely on the variable value ranges. Unfortunately,

4. Conclusions

The philosophical literature on causal capacities has principally asked metaphysical and epistemological questions, rather than representational ones. At the same time, the psychological and machine learning literature on causal capacities has largely focused on the special case of binary causal factors and a binary effect. By generalizing beyond that special case, we thereby obtain a natural, privileged framework for representing causal capacities that independently influence some effect.[13] Moreover, this generalized framework provides further conceptual clarification about causal capacities, as it reveals distinctions (e.g., between the lower bound and the baseline value) that have previously been relatively little-explored in the psychological and machine learning literatures. This mathematical framework also has significant practical and theoretical impacts, as it provides a natural way to unify disparate equations—in particular, the noisy-OR/AND and linear models—that have previously been viewed (in machine learning and cognitive science) as competitors, or at least independent of one another. The widespread use and value of such models is eminently explainable when we understand them as deriving from the

---

cover stories for those experiments almost never explicitly provide value ranges, and we do not know what participants infer about the possible variable values. Anecdotally, though, this type of switching would explain some otherwise puzzling empirical data.

[13] One open question is whether there are also privileged equations for $P(E)$. As a promising first step, we can prove: if there is one generative cause and the initial $P(E)$ is uniform over $[-L, U]$, then the "update" equation $P_{new}(E = e|G = g) = P_{old}(E = U\frac{e - w_g g}{U - w_g g})$ naturally satisfies all of the desiderata provided throughout the paper (including the desired expectation). It is unknown whether other results of this type can be obtained.

same fundamental framework and equation. This privileged framework provides a precise, formal representation that can significantly constrain and inform our attempts to better understand causal capacities.

References

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

———. 1999. *The Dappled World: a Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

———. 2002. "Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward." *The British Journal for the Philosophy of Science* 53: 411–53.

———. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Cheng, Patricia W. 1997. "From Covariation to Causation: a Causal Power Theory." *Psychological Review* 104: 367–405.

Cozman, Fabio G. 2004. "Axiomatizing Noisy-OR." In *Proceedings of the 16th European Conference on Artificial Intelligence*.

Danks, David. 2003. "Equilibria of the Rescorla-Wagner Model." *Journal of Mathematical Psychology* 47: 109–21.

———. 2007. "Causal Learning from Observations and Manipulations." In *Thinking with Data*, edited by Marsha C. Lovett and Priti Shah, 359–388. Mahwah, NJ: Lawrence Erlbaum.

———. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: The MIT Press.

Danks, David, and Clark Glymour. 2001. "Linearity Properties of Bayes Nets with Binary Variables." In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, edited by Jack Breese and Daphne Koller, 98–104. San Francisco: Morgan

Kaufmann.

DeLosh, Edward L., Jerome R. Busemeyer, and Mark A. McDaniel. 1997. “Extrapolation: the Sine Qua Non for Abstraction in Function Learning.” *Journal of Experimental Psychology: Learning, Memory, & Cognition* 23 (4): 968–86.

Glymour, Clark. 1998. “Learning Causes: Psychological Explanations of Causal Explanation.” *Minds and Machines* 8: 39–60.

Good, I. J. 1961. “A Causal Calculus (I).” *British Journal for the Philosophy of Science* 11 (44): 305–18.

Hausman, Daniel M., and James Woodward. 1999. “Independence, Invariance and the Causal Markov Condition.” *The British Journal for the Philosophy of Science* 50: 521–83.

———. 2004. “Modularity and the Causal Markov Assumption: a Restatement.” *The British Journal for the Philosophy of Science* 55 (1): 147–61.

Heckerman, David, and John S. Breese. 1994. “A New Look at Causal Independence.” In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, 286–92. Morgan Kaufmann.

———. 1996. “Causal Independence for Probability Assessment and Inference Using Bayesian Networks.” *IEEE Transactions on Systems, Man, and Cybernetics: Part a Systems and Humans* 26 (6): 826–31.

Heil, John. 2005. *From an Ontological Point of View*. New York: Oxford University Press.

Holyoak, Keith J., and Patricia W. Cheng. 2011. “Causal Learning and Inference as a Rational Process: the New Synthesis.” *Annual Review of Psychology* 62: 135–63.

Hyttinen, Antti, Frederick Eberhardt, and Patrik O. Hoyer. 2011. “Noisy-or Models with Latent Confounding.” In *Proceedings of the 27th Conference on Uncertainty in Artificial*

Intelligence.

———. 2012. "Learning Linear Cyclic Causal Models with Latent Variables." *Journal of Machine Learning Research* 13: 3387–3439.

Kalish, Michael L., Thomas L. Griffiths, and Stephan Lewandowsky. 2007. "Iterated Learning: Intergenerational Knowledge Transmission Reveals Inductive Biases." *Psychonomic Bulletin & Review* 14: 288–294.

Kim, Jin H., and Judea Pearl. 1983. "A Computational Model for Causal and Diagnostic Reasoning in Inference Systems." In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, 190–93. San Francisco: Morgan Kaufmann.

Lucas, Peter J. F. 2005. "Bayesian Network Modeling Through Qualitative Patterns." *Artificial Intelligence* 163: 233–63.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67 (1): 1–25.

Martin, C. B. 2008. *The Mind in Nature*. Oxford: Oxford University Press.

McDaniel, Mark A., and Jerome R. Busemeyer. 2005. "The Conceptual Basis of Function Learning and Extrapolation: Comparison of Rule-Based and Associative-Based Models." *Psychonomic Bulletin & Review* 12 (1): 24–42.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible iInference*. San Francisco: Morgan Kaufmann Publishers.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: The MIT Press.

Srinivas, Sampath. 1993. "A Generalization of the Noisy-OR Model." In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence*, 208–15.

Bridge Laws and the Psycho-Neural Interface *†

Marco J. Nathan* and Guillermo Del Pinal**

*Department of Philosophy, University of Denver
**Department of Philosophy, Columbia University & Mercator Memory Research Group, Ruhr Universität Bochum

August 16, 2014

Abstract

Recent advancements in the brain sciences have enabled researchers to determine, with increasing accuracy, patterns and locations of neural activation associated with various psychological functions. These techniques have revived a longstanding debate regarding the relation between the mind and the brain: while many authors now claim that neuroscientific data can be used to advance our theories of higher cognition, others defend the so-called 'autonomy' of psychology. Settling this significant question requires understanding the nature of the *bridge laws* used at the psycho-neural interface. While these laws have been the topic of extensive discussion, such debates have mostly focused on a particular type of link: *reductive laws*. Reductive laws are problematic: they face notorious philosophical objections and they are too scarce to substantiate current research at the interface of psychology and neuroscience. The aim of this article is to provide a systematic analysis of a different kind of bridge laws—*associative laws*—which play a central, albeit often overlooked, role in scientific practice.

1 Introduction

In a now classic paper, Jerry Fodor (1974, p. 97) questioned the evidence for theoretical reductionism by noting that "the development of science has witnessed the proliferation of specialized disciplines at least as often as it has witnessed their reduction to physics, so the widespread enthusiasm for reduction can hardly be a mere induction over its past successes." Four decades later,

*Both authors contributed equally to this work.

†We are grateful to Bruce Pennington and Kateri McRae for constructive comments on various versions of this essay.

Fodor's assessment remains accurate; indeed, it has been reinforced. Rather than being progressively reduced to physics, the special sciences have sprawled into a number of burgeoning subfields. Yet, at the same time, we have also witnessed the rise of *interdisciplinary* studies. If, as Fodor holds, the special sciences are relatively 'autonomous,' what explains the recent proliferation of fields such as neurolinguistics, moral psychology, and neuroeconomics?

The relation between different scientific fields has been extensively debated in philosophy and the particular case of psychology and neuroscience has gathered enormous attention. As reported in Bourget and Chalmers (2013), the dominant position is now *non-reductive physicalism*—the thesis that, although mental states are realized by brain states, mental kinds cannot, in general, be reduced to neural kinds. As we shall discuss below, this position fails to address an important issue, namely, why studying the brain can inform our understanding of the mind. The failure to provide an answer to this question is especially troublesome given the current trend in cognitive neuroscience, where advancements in neuroimaging have begun to affect theories of higher cognition, such as language processing and decision making (Gazzaniga 2009; Mather et al. 2013; Glimcher and Fehr 2014). If theorists are right that the mapping of mental kinds onto neural kinds is too problematic to substantiate any meaningful interaction at this interface, is neuroscience simply promising something that cannot be achieved? Or does the constant use of neural data in fields such as neurolinguistics and neuroeconomics, show that philosophical critique misunderstands the relation between cognitive and neural levels?

In this article, we argue that the tension between meta-theory and scientific practice stems from the failure to distinguish between different types of bridge laws, that is, principles that link kinds across domains. On the one hand, theorists have generally been concerned with *reductive* laws, which are indeed problematic. On the other hand, bridge laws currently employed in cognitive neuroscience are not reductive; they are *associative* statements that are categorically distinct from the contingent type-identities typically employed in derivational reduction and other more recent reductive approaches. The aim of this essay is to provide an account of associative bridge laws. Despite their widespread use in neuropsychology, these links have never been systematically discussed. We begin by introducing the role of type-identities in traditional models of derivational reduction and rehearse some well-known problems (§2). Next, we illustrate how bridge laws are employed in neuroscientific studies of higher-cognition (§3) and elucidate the main differences between reductive and associative bridge laws (§4). We conclude by presenting some implications of our analysis for extant debates in the philosophy of mind and science (§5).

2 Bridge Laws in Theory Reduction

Philosophers of science originally became interested in bridge laws because of their central role in theory reduction. In what became a *locus classicus*, Nagel (1961) characterized reduction as a deductive derivation of the laws of a reduced

theory S from the laws of a reducing theory P. Such derivation requires that the natural kinds of S be expressed in terms of the natural kinds of P.[1] For instance, suppose that we want to show that law $L_S : S_1x \rightarrow S_2x$, expressed in the language of theory S, can be reduced to (that is, derived from) law $L_P : P_1x \rightarrow P_2x$, expressed in the language of theory P.[2] What we need is a series of bridge laws that translate the relevant S-predicates into P-predicates:

(B_1) $S_1x \leftrightarrow P_1x$

(B_2) $S_2x \leftrightarrow P_2x$

How should the '$\leftrightarrow$' connective be interpreted in these reductive bridge laws? Fodor (1974) makes a number of important points. First, $\leftrightarrow$ must be *transitive*: if kind S_1 is reduced to T_1, and in turn, T_1 is reduced to P_1, then S_1 is thereby reduced to P_1. Second, $\leftrightarrow$ cannot be read as 'causes,' for causal relations tend to be *asymmetric*—causes bring about their effects, but effects generally do not bring about their causes—whereas bridge laws tend to be *symmetric*: if an S_1-event is a P_1-event, then a P_1-event is also an S_1-event. Given these two features, bridge laws are most naturally interpreted as expressing *contingent event identities*. Thus understood, B_1 can be read as stating that S_1 is *type-identical* to P_1. As Fodor notes, reductive bridge laws express a stronger position than *token physicalism*, the view that all events that fall under the laws of some special science are physical events. Statements such as B_1 and B_2 presuppose *type physicalism*, according to which every kind that figures in the laws of a science is type-identical to a physical kind.[3]

The well-known problem with type-physicalism is that natural kinds seldom correspond neatly across levels. Although one could make a case that heat is reducible to mean molecular kinetic energy, or action-potentials to nerve impulses, the reigning consensus in philosophy of science is that there are too few contingent event identities to make derivational reduction a plausible inter-theoretic model (Horst 2007). In most cases, there seem to be no physical, chemical, or macromolecular kinds that correspond to biological, psychological or economic kinds in the manner required by the reductionist scheme. This, simply put, is the *multiple-realizability argument* against the classical model of derivational reduction (Putnam 1967; Fodor 1974). The basic idea is that instead of laws such as B_1 and B_2, what we usually find are linking laws such as B_3, which capture the instantiation of higher-level kinds in a variety of lower-level states:

(B_3) $S_1x \leftrightarrow P_1x \vee \ldots \vee P_nx$

[1] In what follows we shall not enter the longstanding metaphysical debate on the notion of *natural kind*. For present purposes, we treat natural kinds as predicates that fall under the laws or generalizations of a (branch of) science (Fodor 1974).

[2] For the sake of simplicity, we assume that the languages of the two theories do not overlap, i.e., the natural kinds of S do not belong to P, and vice versa.

[3] To be clear, our focus here is not on *physicalism per se*; the relevant claim is whether the kinds of one science can be reduced to the kinds of another more 'fundamental' science, not necessarily to physics.

The demise of derivational reduction had a deep and lasting effect on the conceptualization of the psycho-neural interface. Despite its problems, the Nagelian model provided a clear account of how neural data could, at least in principle, inform theories of higher cognition. To illustrate, suppose we want to know whether some psychological kind C is engaged in task T, as we often do when testing competing cognitive-level hypotheses. If we had a bridge law which maps C onto a neural kind N, we could infer the presence (or absence) of C in T from neural evidence of the presence (or absence) of N. Hence, the reductive model suggests a specific goal for cognitive neuropsychology, namely, to look for neural-level implementations of psychological processes. The failure of Nagelian reduction, however, implies that this account of the psycho-neural interface is misguided or, at best, overly simplistic.

In response to the multiple-realizability argument, philosophers pursued two alternative routes. Some reacted by developing reductive accounts that, allegedly, do not require problematic bridge laws (Hooker 1981; Bickle 1998; Kim 1999, 2005). However, it has been persuasively argued that any form of *bona fide* reductionism requires some kind of bridge laws (Marras 2002; Fazekas 2009). Following a different path, many philosophers of mind embraced an antireductionist or functionalist approach, according to which mental states are individuated by their causal roles, independently of their physical realization (Fodor 1974, 1997). While this move besets the problems raised by multiple realizability, it fails to explain how, if cognitive kinds are not type-identical to neural kinds, neural data can bear on the study of cognition.

Part of the problem with the extant debate, we surmise, is that reductionists and antireductionists alike share an overly restrictive view of the psycho-neural interface. Researchers belonging to both camps often talk as if the only potential contributions of neuroscience to psychology are:

(i) To establish *correlations* between cognitive- and neural-level events, i.e., to find the brain locations *where* particular mental functions are computed.

(ii) To discover the neural-level mechanisms that *compute* cognitive processes, i.e., to establish *how* the brain actually computes specific mental functions.

Let us begin by focusing on (ii), the more substantial and ambitious endeavor. Reductionists tend to stress the remarkable successes in discovering neural mechanisms of sensory systems, such as early vision, pain, taste, and other basic sensations (Bickle 2003; Kim 2006). Antireductionists, in contrast, rightly emphasize that comparable achievements cannot be claimed for language processing, decision making, and other functions of higher cognition. It is unsurprising, then, that many researchers deem the pursuit of project (ii) hopeless (Fodor 1999) or, at best, drastically premature (Gallistel 2009; Coltheart 2013), at least when applied to central cognitive systems. On the traditional view of the interface based on (i) and (ii) this skepticism is reasonable. Although perceptual functions are potentially multiply realizable, empirical research reveals that they are implemented by relatively modular and localized neural structures, widely shared across individuals and species. In contrast, systems of

higher cognition are implemented by relatively flexible, distributed, and non-modular neural structures. Thus, in the case of higher cognition, the pursuit of project (ii) is jeopardized by multiple realizability and the lack of explanatory reductions. But, note, if (ii) is hopeless, (i) becomes pointless, for seeking mind-brain correlations that do not contribute to an explanation of *how* neural mechanisms compute cognitive functions becomes a mere vindication of *token physicalism*. In short, from this perspective, project (ii) becomes unrealistic and project (i), by itself, can hardly advance studies of higher cognition.[4]

Despite this bleak picture, it is undeniable that interdisciplinary fields at the psycho-neural interface, such as neurolinguistics and neuroeconomics, have recently achieved remarkable success, often by using neural-level data to advance cognitive level theories.[5] Neither reductive nor antireductive models can appropriately account for this. Still, these studies presuppose that it *is* possible to map the cognitive level onto the neural level for, otherwise, how can neural data be used to bear on cognitive-level theorizing? In order to account for the success of these interdisciplinary studies, we need a novel account of bridge laws that takes seriously their non-reductive character. To explore the nature of these links, we shall focus on one of the main techniques which scientists use to make neural data and theories bear on cognitive level hypotheses: *reverse inference*.

3 Bridge Laws and Reverse Inferences

In order to discriminate between competing cognitive hypotheses, neuroscientists often 'reverse infer' the engagement of a cognitive state or process, in a given task, from particular locations or patterns of brain activation (Henson 2005; Poldrack 2006; Del Pinal and Nathan 2013; Hutzler 2013; Machery 2013). These *reverse inferences* presuppose the availability of bridge laws; yet, contrary to a widespread assumption, the required links are not reductive, they are what we call *associative bridge laws*. In this section, we examine the role of bridge laws in two kinds of inferences employed in neuroimaging studies: *location-based* and *pattern-based reverse inferences*. More specifically, we focus on studies of decision-making—a paradigmatic domain of higher-cognition—aimed at discriminating between the processes which underlie behavioral generalizations.

To begin, consider the following psychological generalizations, somewhat simplified for the sake of illustration, where s ranges over 'normal' adults:

[4]Those familiar with this debate will no doubt have seen various objections along these lines. For instance, the picture of the psycho-neural interface assumed in the following quotes is clearly constrained by (i) and (ii). "If the mind happens in space at all, it happens somewhere north of the neck. What exactly turns on knowing how far north?" (Fodor 1999). "Finding a cell that recognizes one's grandmother does not tell you very much more than you started with: after all, you know you can recognize your grandmother. What is needed is an answer to how you, or a cell [...] does it" (Mayhew 1983, cited in Coltheart (2013)).

[5]To appreciate the magnitude of this growth, consider that in 2009, when the first canonical textbook was published (Glimcher et al. 2009), courses and research on neuroeconomics were regularly taught and pursued in just handful of economics and psychology departments. By the time the second edition appeared, just four years later (Glimcher and Fehr 2014), over one hundred institutions regularly taught and pursued research in neuroeconomics.

(G_M) If s is faced with the option of performing an action a that will result in the death of fewer people than would die if s were not to perform a, s will choose a unless doing so requires using a person directly as a means.

(G_N) A set E contains some items that are new to s and others that s has previously encountered. If s is randomly presented with item $e \in E$ and has to decide whether she has previously encountered e, s can reliably distinguish between old and new items.

G_M and G_N can be refined in various ways, but neither is particularly original nor controversial. Both capture distinctive capacities of higher-cognition which are in need of explanation. We shall refer to the level at which we isolate these types of psychological generalizations as *Marr-level 1*.[6]

Given a Marr-level 1 generalization, one can then explore the underlying cognitive processes: such conjectures are usually referred to as *Marr-level 2 hypotheses*. First, consider two competing explanations of G_M:

(M) In moral decision making, subjects generally follow consequentialist rules. However, in cases which involve using another person directly as a means, consequentialist rules are overridden by *negative emotions*.

(M^*) In moral decision making, subjects generally follow consequentialist rules. However, in cases which involve using another person directly as a means, consequentialist rules are overridden by *deontological rules*.

Note that M and M^* are very different explanations of G_M. Whereas M explains the behavioral pattern as a conflict between rules and emotions, M^* explains the same pattern as a conflict between consequentialist and deontological rules. In short, while M posits a conflict between rules and emotions, M^* posits a conflict between different types of rules. Next, consider two competing explanations of G_N, recently advanced in episodic memory research:

(N) Recognition decisions are based on two processes which draw on two distinct sources of information: *recollection* of specific details and non-specific feelings of *familiarity*. Recollection is used by default but, when such information is unavailable, subjects employ familiarity.

(N^*) Recognition decisions are based on two processes which draw on two distinct sources of information: *recollection* of specific details and non-specific feelings of *familiarity*. However, neither is the default process: the source of information employed depends on *specific contextual cues*.

[6] In an influential discussion, Marr (1982) argued that information-processing systems should be investigated at three complementary levels. Hypotheses at Marr-level 1 pose the computational problem: they state the task computed by the system. Hypotheses at Marr-level-2 state the algorithm used to compute Marr-level 1 functions: they specify the basic representations and operations of the system. Finally, hypotheses at Marr-level 3 specify how Marr-level 2 algorithms are implemented in the brain: they purport to explain *how* these basic representations and operations are realized at the neural level.

While N and N^* agree on the basic components underlying recognition decisions, they posit different interactions. According to N, subjects generally use recollection information to decide whether items are old, and only rely on intuitions of familiarity when such information is unavailable. In contrast, N^* predicts that certain contextual cues will induce subjects to make familiarity-based recognition decisions even if recollection information is available.

M-M^* and N-N^* are competing Marr-level 2 hypotheses about the cognitive processes which underlie some Marr-level 1 generalization. To adjudicate between them, researchers use reverse inferences, which require two preliminary steps. First, the competing processes must be functionally decomposed, for entire processes such as M and M^* are too coarse-grained to be directly mapped onto patterns or regions of neural activation. Next, the subcomponents of the competing processes for which there are bridge laws must be identified. To illustrate, let us assume that, in task T, cognitive process M posits the engagement of subprocess m_1, whereas M^* posits the engagement of subprocess m_1^*, and that $m_1 \neq m_1^*$. Further, suppose that we have the following bridge laws connecting m_1 and m_1^* with regions or patterns of neural activation n_1 and n_1^*:

(A_1) $m_1 \otimes n_1$

(A_2) $m_1^* \otimes n_1^*$

Note that '$\otimes$' is different from the '$\leftrightarrow$' connective figuring in reductive bridge laws. We shall discuss the basic properties of such relation in §4 below. The important point here is simply that '$\otimes$' stands for an associative relation that allows one to reliably infer the presence of one relata from the other.

To illustrate the application of statements such as A_1 and A_2, consider some bridge laws used to discriminate between M and M^*. Assume that m_1 stands for processes involving negative emotions such as fear, and that m_1^* stands for ruled-based processes such as following simple instructions. Researchers have established a close connection between processes involving negative emotions and activation in certain neural regions such as the amygdala and the ventromedial prefrontal cortex (VMPFC).[7] This connection is captured by A_1. Researchers have also established a connection between rule-based and controlled reasoning and activation in the dorsolateral prefrontal cortex (DLPFC).[8] A_2 captures this connection by associating m_1^* with activation in the DLPFC.

[7] In general, the amygdala is critically involved in conditioned and unconditioned fear response in animals, including humans. For example, patients with selective damage to the amygdala show no physiological response to a previously fear-conditioned stimulus, although they can explicitly remember the conditioning experience (Kandel et al. 2013, Ch. 48).

[8] Miller and Cohen (2001) present several studies that support the key role of the DLPFC in cognitive control and rule-guided processes. A relevant set of experiments are based on the famous Stroop task, in which subjects are instructed to name the color of the ink of words as they appear on a screen. Famously, reaction times and error rates increase dramatically when subjects read color-terms that differ from the color of their ink. Miller and Cohen present imaging studies which show that, in the misleading cases, subjects who manage to follow the correct rule and name the word's ink color showed increased activation in DLPFC, compared to subjects who fail the task.

Given A_1 and A_2, one can devise neuroimaging experiments to discriminate between M and M^*. For example, Greene and colleagues (2001) scanned subjects making moral decisions in two sets of tasks that involve choosing whether to sacrifice one innocent person to save five, as in the famous trolley problems. The relevant difference is that in one set of tasks all the choices that would save five people involve using another person directly as a means (*personal cases*), whereas in the other set subjects can save five by sacrificing one indirectly, that is, without using the person as a means (*impersonal cases*).[9] Greene and colleagues found that, relative to impersonal cases—and to structurally analogous non-moral control tasks—personal cases result in differential activation of the amygdala and VMPFC, and less activation of DLPFC. Given that A_1 associates amygdala activation with negative emotions, and that A_2 associates DLPFC activation with rule-based and controlled reasoning, this finding favors M over M^*. This is because, according to M, in personal cases, decisions not to sacrifice one person to save five are based on negative emotions. In addition, M predicts that areas involved in rule-based reasoning should be more active in impersonal compared to personal cases. In contrast, M^* incorrectly predicts that personal and impersonal cases should engage rule-based areas equally, since both cases involve applying different types of rules.

Critics of the relevance of neuroimaging experiments for psychology often assume—more or less explicitly—that all bridge laws currently employed in reverse inferences associate cognitive processes to *locations* of neural activation. However, as we shall discuss below, this is a mistake: in some cases, the relevant bridge laws map cognitive states or processes to particular *patterns* of neural activation. Indeed, pattern-based inferences, which are rapidly becoming one of the main ways of studying cognition, have significant implications for the psychoneural interface. A powerful example is provided by recent studies relevant to the recognition hypotheses N and N^*, to which we now turn.

In pattern-based recognition studies, 'pattern classifiers' are trained to determine the multi-voxel patterns associated with recollection processes and familiarity processes. Specifically, classifiers are trained in tasks where experimenters can control which cognitive process is engaged. For instance, in one experiment, which will serve as our main example, subjects were exposed to singular and plural words such as 'shoe' and 'shoes' (Norman et al. 2009). These subjects were then scanned while performing recognition tasks involving previously examined items (e.g., a shoe) and unrelated lures (e.g., a bicycle). The recognition tasks are divided in two sets: *recollection blocks* and *familiarity blocks*. In recollection blocks, subjects are instructed to recall specific details of the mental image formed during the study phase, and to only answer 'yes' if they are successful. In contrast, in familiarity blocks subjects are instructed to answer 'yes' if the word is familiar and to ignore any details they might recollect from the study phase. After training, classifiers can determine whether some multi-voxel

[9] In the classic version of the trolley problem, personal cases are exemplified by the 'footbridge' scenario, where five people are saved by throwing a corpulent person on the track. Impersonal cases are exemplified by the 'switch' scenario, where five people are saved by pulling a lever that diverts the trolley onto a parallel track where it will kill a single person.

pattern of neural activation is an instance of recollection or familiarity. What makes this method especially interesting is that the reliability of the classifiers can be established within the experiment itself. This can be done by saving a subset of the recollection and familiarity blocks for later testing (so they are not used at the training stage), and then determining the rate at which the classifier correctly categorizes the corresponding neural patterns. This part of the study, in which experimenters control which process is engaged, establishes the bridge laws that will then be used in reverse inferences.

Having established the relevant bridge laws which map recollection and familiarity onto multi-voxel patterns, one can then test competing hypotheses N and N^* regarding the dynamics underlying recognition-decisions in cases where the engagement of the sub-processes cannot be directly controlled. For example, in a second phase of the study, subjects were scanned while trying to determine whether some word is old or new, while being exposed to previously studied items ('shoe' and 'ball'), unrelated lures ('horse' and 'box'), and previously unstudied switch-plurality lures ('balls'). Experimenters then examined the subset of the items for which subjects made correct positive recognition decisions. Note that these are cases where both recollection and familiarity information was available to subjects. Hence, according to hypothesis N, the classifier should categorize the corresponding voxel patterns as recollection patterns (since this is the default). In contrast, N^* predicts that the classification should be more variable, involving—at least in some cases—familiarity patterns. Experimental results support N^* over N: when both types of information are available, various contextual cues determine whether subjects use familiarity or recollection as the basis of their recognition decision (Norman et al. 2009).

4 Associative Bridge Laws

The previous examination of reverse inferences allowed us to place associative bridge laws such as A_1 and A_2 in their context of use. The aim of this section is to make explicit the characteristic features of these linking statements. As we shall see, unlike their reductive counterparts, associative bridge laws are *probabilistic* and *context-sensitive* relations that do *not identify* their relata, either at the type-level or at the token-level.

4.1 Probabilities

The first main feature of associative bridge laws is their *probabilistic* nature. To clarify, consider a recent debate about the 'selectivity' of brain regions and reverse arguments. Several critics have emphasized that the success of a reverse argument depends on the degree of selectivity of the relevant brain regions (Uttal 2002; Ross 2008; Phelps 2009; Anderson 2010; Coltheart 2013). Suppose that some bridge law maps neural activation in n_1 onto the engagement of cognitive process m_1. According to critics, this linkage allows one to legitimately reverse infer the engagement of m_1 from the activation of n_1 only provided that region

n_1 activates for the cognitive process of interest, in this case n_1, *and no other*. This is because, the objection runs, if n_1 also activates when m_2, m_3, and m_4 are engaged, one *cannot* reverse infer to m_1 merely on the neural evidence of n_1 activation. The problem is that there is widespread consensus among cognitive neuroscientists that very few brain regions are *maximally selective* in the sense just described. From this perspective, then, it looks like most reverse inferences are actually invalid, as they rely on an unjustified maximal selectivity.

This is a substantial worry that ought to be addressed with care. First, note that while few brain regions are indeed maximally selective, most brain regions are not mapped onto cognitive functions by a single bridge law. Most brain regions are covered by *multiple* bridge laws which associate them with a variety of cognitive functions. Consequently, when we reverse infer the engagement of a cognitive function from the activation of a neural region, the inference falls short of absolute certainty. Confidence that one has identified the correct bridge law is a matter of degree, which is determined by the conditional probability that cognitive process m_1 is engaged, given activation in n_1.[10] As an illustration, consider, again, the example of moral decision making. As neuroscientists know, the amygdala is also activated by processes that are not related to negative emotions in any obvious way; consequently, amygdala activation does not deductively entail the engagement of fear or similar emotions. However, it does not follow that inferences from amygdala activation to the presence of negative emotions are invalid; what follows is simply that such inferences are *inductive* or *probabilistic*. The case of the amygdala is not the exception, it is the norm: most brain regions are associated with various cognitive processes or states. Furthermore, this point is not restricted to location-based inferences, but also applies to pattern-based ones. The multi-voxel patterns are, at best, a reliable guide for inferring (*via* bridge laws) the engagement of the associated cognitive state or process.

With all of this in mind, we can now turn to an influential critique of the probabilistic nature of reverse inferences. Several authors have argued that, since the application of a given bridge law in some task is determined by a conditional probability, most interesting reverse inferences turn out to be unacceptably weak (Miller 2008; Phelps 2009; Legrenzi and Umilta 2011). This objection underlies many skeptical claims about the use of reverse inferences and has led to the explicit suggestion that genuine progress at the psycho-neural interface requires reductionist bridge laws (Ross 2008; Anderson 2010). No doubt, in some cases, such accusations are justified: some proposed reverse inferences

[10]This conditional probability is determined by the following straightforward application of Bayes' theorem:

$$P(m_1|n_1) = \frac{P(n_1|m_1)P(m_1)}{P(n_1|m_1)P(m_1) + P(n_1|\neg m_1)P(\neg m_1)} \quad (1)$$

Note that the prior $P(m_1)$ is conditioned on the task used in the reverse argument. Importantly, Equation (1) shows that the degree of belief in a reverse inference depends not only the prior $P(m_1)$ but also on the selectivity of the neural response—i.e., on the ratio of the process-specific activation, $P(n_1|m_1)$, to the overall likelihood of activation in that area across all tasks which do not involve m_1, i.e., $P(n_1|\neg m_1)$.

are indeed questionable, to say the least. Yet, this observation falls short of a general critique, for the significance of the lack of (maximal) selectivity on the validity of reverse inferences has been substantially exaggerated. This is because critics often overlook another important characteristic of associative bridge laws, namely, their *context sensitivity*.

4.2 Context-Sensitivity

In an influential article, Poldrack (2006) noted that the conditional probability that a cognitive state m_1 is associated to a neural state or process n_1 should be determined *relative to a particular task*. However, to avoid unnecessary complications, Poldrack intentionally ignored this task-relativity in the rest of his analysis. This deliberate omission, however, had the unfortunate consequence that several ensuing discussions also ignored the task-relativity of bridge laws in reverse inferences. This resulted in a misleading objection.

Consider the selectivity of the amygdala, which plays a central role in several studies in neuroethics and neuroeconomics. Although the amygdala is typically involved in processes involving fear and other negative emotions, it is also involved in many other cognitive processes that are usually unmentioned in studies such as Greene et al. (2001). Such processes include the perception of odor intensity, sexually arousing stimuli, and trust from faces (Phelps 2006; Lindquist et al. 2012), as well as the processing of faces from other races, and the perception of biological motion and sharp contours (Phelps 2009). It has also been claimed that the main function of the amygdala is to process novel or emotionally salient stimuli—not fear-related stimuli *per se* (Lindquist et al. 2012). Based on these considerations, Phelps (2009) argues that amygdala activation in a given psychological task could signal the engagement of *any* of these cognitive processes. Consequently, reverse inferences such as the ones used by Greene and colleagues overestimate the conditional probability that negative emotions are engaged, given amygdala activation.

What Phelps and other critics (e.g., Klein 2011) overlook is that the probability that a particular bridge law applies, given the activation of a brain region, should be determined relative to relevant features of the context invoked by the reverse argument. Specifically, in the case under consideration, the success of the reverse argument does *not* depend on the assumption that we can reliably infer the engagement of negative emotions from differential activation in the amygdala. What the argument requires is that the engagement of negative emotions can be inferred from the pattern of neural activation observed *in the particular task under consideration*.[11] In other words, the inference is from differential amygdala-activation *in personal scenarios* to the engagement of negative emotions. Once the inference is framed in these terms, we can see that most other

[11] For a discussion of task-relativity in reverse inferences, see Hutzler (2013) and Del Pinal and Nathan (2013). In a related discussion, Machery (2013) defends the relativity of the cognitive-level *hypotheses* being tested. Despite significant differences, here we can treat all these approaches on a par, for they address the 'lack of selectivity' objection by emphasizing the inherent relativity of reverse inferences.

cognitive processes that also involve the amygdala are not plausible explanations for such differential activation, and can thus be ruled out. Consider, for instance, the tasks used by Greene and colleagues (2001). Personal cases do not differ from impersonal ones with respect to stimuli related to odor, facial-processing, sexuality, sharp-contours, or the comparative novelty of the tasks. Hence, relative to personal cases, the conditional probability of the engagement of negative emotions, given amygdala activation, is significantly higher than is suggested by the objection presented above.[12]

The critiques against reverse inference based on lack of selectivity—which are typically raised against location-based inferences—become even less persuasive when directed against pattern-based inferences. Yet, we should explicitly stress that, just like location-based ones, pattern-based inferences are also context-sensitive. For instance, the recognition experiments discussed in the previous section employ bridge laws that associate particular multi-voxel patterns with recollection and familiarity processes. In tasks that contrast recollection- and familiarity-based recognition judgments, each set of multi-voxel patterns can be used by a classifier to reliably identify instances in which recollection or familiarity are engaged. However, these inferences are especially useful because, as noted in §3, the reliability of the classifier can be established, directly and precisely, in an experimental setting. In general, pattern-based inferences are more reliable than location-based ones; still, both are context-sensitive in essentially the same way.

4.3 Non-Identity

Unlike their reductive counterparts, associative bridge laws do not presuppose any kind of identity—*a priori*, *a posteriori*, *necessary*, or *contingent*. To wit, in the moral decision making case, the bridge law mapping amygdala activation to the engagement of negative emotions presupposes neither the type-identity nor the token-identity of these two events. As we saw, the amygdala is differentially activated by a variety of cognitive processes that have little or nothing to do with negative emotions, and it might turn out that some unambiguously fear-or-distress-related processes are not accompanied by increased amygdala activation. We should make it very clear that we are not recommending any departure from token-physicalism. Our point is simply that associative bridge laws are so metaphysically uncommitted that they would also be consistent with violations of token-physicalism.

A similar point applies to pattern-based inferences. Bridge laws used in the recognition case do not presuppose that recollection or familiarity processes are (type- or token-) identical to their associated multi-voxel patterns. For one

[12] We surmise that the task relativity of reverse inferences is systematically overlooked because methodological discussions (e.g., Poldrack 2006; Phelps 2006) often consider only arbitrary 'empty' tasks which do not eliminate any processing possibilities (that is, any bridge laws) for the brain region of interest. Hence, reverse inferences seem intuitively weak. However, once we consider the tasks relevant to each reverse inference, we can eliminate some subset of bridge laws which cover the brain regions of interest, thereby increasing their strength.

thing, the patterns are only highly reliable—but not infallible—indicators of the corresponding processes. Furthermore, and more importantly, even if we had perfect correlations, multi-voxel patterns are not plausible candidates for such identities. Voxel patterns are representations that average over the activation of thousands of neurons, but do not specify the actual neural mechanisms that compute cognitive-level processes. This, of course, is not to say that the possibility of a type-identity can be ruled out *a priori*: one might believe that, eventually, the neural mechanisms that carry out, say, recollection processes will be identified. However, this reduction is neither required nor presupposed by the use of pattern-based inferences to discriminate among competing hypotheses of the processes underlying recognition tasks.

To appreciate the main features of associative bridge laws, it is useful to contrast them with various recent attempts that deal with multiple realizability by weakening Nagelian bridge laws. David Lewis (1969) famously argued that reductive type-identities are not meant to hold across the board. On his view, the bridge laws reducing mental states to brain states are implicitly restricted to a specific domain. For example, while pain *tout court* cannot be reduced to a single brain state, human pain, octopus pain, martian pain, etc. can each be reduced to a different type of brain state. Lewis' argument has been subsequently developed and refined by various philosophers (Hooker 1981; Enç 1983; Churchland 1986; Kim 1992) all of whom pointed out the conditional nature of virtually all contingent event identities.[13] Whether or not the context-relativization of bridge laws is ultimately successful (which has been the subject of heated discussion), it is irrelevant to the present approach. Associative bridge laws do not require restricted conditional identities of any kind. This is especially evident in the case of pattern-based inferences: the particular voxel patterns used to infer the engagement of each sub-type of recognition process—that is, the bridge laws—are not even stable across individuals, let alone all human beings, and can only be used reliably in specific experimental contexts. In the experiments considered above, the voxel patterns were used to infer the engagement of familiarity or recollection in a task where these processes were the only unknown variables. If a third task (say, a face-recognition process) were added, the pattern-classifier would have to be re-trained. In this case, there would be no guarantee that the patterns that were previously associated with familiarity and recollection could still be used, in the new experimental settings, to reliably predict those same processes.

For similar reasons, associative bridge laws should also be distinguished from recent attempts to weaken Nagelian bridge laws by replacing type-identity with a condition of *connectability* based on co-referentiality. Klein (2009) argues that a higher-level science S is N-connectable to a lower-level science S' if and only if S' has the resources to introduce new terms, in its own vocabulary, which are *co-referential* with the predicates of S that are absent in S'. Determining

[13] To cite a textbook example, the standard identification of temperature with mean molecular kinetic energy in classical equilibrium thermodynamics is left completely unscathed, the arguments runs, by the observation that temperature is differently realized in gases, solids, vacuums, and other mediums.

the co-referentiality of terms is a substantial endeavor that, however, we can set aside. The important point is that whether or not terms such as 'amygdala activation' and 'fear' are co-referential—and there seems to be no reason to assume that they are, given that one is often found without the other—is irrelevant to our account, for the co-referentiality of terms is not a precondition for their successful employment in reverse inferences.

In sum, the bridge laws which figure in location-and pattern-based reverse arguments do not assume any kind of identity between neural and cognitive states or processes. In order to play a role at the psycho-neural interface, associative bridge laws only need to allow us to reliably (reverse) infer, in certain experimentally controlled settings, the engagement of a cognitive state or process from particular locations or patterns of neural activation.

5 Implications

In the previous section, we analyzed the characteristic features of associative bridge laws by drawing on the way they are employed in scientific practice and contrasting them with their reductive counterparts. We now turn to their implications for various ongoing debates about inter-level relations in philosophy of mind and science. Specifically, we begin by discussing functional locationism and multiple realizability. We conclude by revisiting the traditional intrerpretation of Marr-levels and its relation to the alleged 'automony' of psychology.

5.1 Avoiding Radical Locationism

Many scholars, including prominent scientists and philosophers, argue that cognitive neuroscientists assume an unreasonably strong version of *functional locationism* (Van Orden and Paap 1997; Fodor 1999; Uttal 2001; Coltheart 2013; Satel and Lilienfeld 2013). Some have gone as far as labeling current cognitive neuroscience the 'new phrenology' (Uttal 2002). This critique often presupposes a reductive model of inter-level relations at the psycho-neural interface. To wit, if one combines the assumptions that said links are reductive and that most reverse inferences are still grounded in lesion studies and location-based neural data, it becomes reasonable to conclude that cognitive neuropsychologists are in the business of type-identifying cognitive functions with neural locations, blatantly ignoring multiple realizability and the failures of derivational reduction. While the charge of excessive functional locationism is sometimes warranted, it does not apply to properly conducted reverse inferences (Del Pinal and Nathan 2013). Furthermore, it does not reflect the current trend in cognitive neuroscience, at least if the increasing importance of pattern-based inferences is a reliable indicator (Poldrack 2008, 2011).

As illustrated by our examples, most reverse arguments do not associate the engagement of entire cognitive processes with specific locations of neural activation. The general strategy is to decompose the competing processes into their subcomponents and to consider those subcomponents that can be mapped, *via*

bridge laws, to neural locations or patterns, from which we can reliably reverse infer the engagement of one of the cognitive processes, relative to a specific task. In the moral case, only one of the competing processes predicted the engagement of negative emotions in personal tasks, which is why differential amygdala-activation provided evidence in favor of M over M^*. The point to stress is that, for the argument to go through, one need not assume the functional localization of the entire moral decision-making processes. Pattern-based inferences are even less plausible targets for the charge of unjustified functional locationism. Classifiers use multi-voxel patterns to infer the engagement of recollection or familiarity in particular recognition tasks. Note that classifiers are given no location-related information, which allows for the set of patterns assigned to, say, recollection to be implemented in different neural locations. Interestingly, recent studies suggest that key components of recognition processes are, indeed functionally localized (Norman et al. 2010). Yet the reverse inference does not presuppose any link between neural patterns and locations of activation. To be sure, there remain several controversial issues regarding the foundations of cognitive neuropsychology, including the substantial question of how to formalize the context- or hypothesis-relativity of reverse inferences (Del Pinal and Nathan 2013; Hutzler 2013; Machery 2013). Yet, the wholesale dismissal of the entire cognitive neuropsychology of higher cognition as a 'sophisticated new phrenology in disguise' does not withstand serious scrutiny.

5.2 Accommodating Multiple Realizability

As discussed in §2, the natural kinds of a 'higher' science cannot, in general, be reduced to kinds of a 'lower' science because natural kinds seldom correspond across domains in the way required by reductive bridge laws. A complete assessment of multiple realizability and reduction lies beyond the scope of this article. Our point is simply that multiple-realizability, coupled with a reductive conception of bridge laws, generates serious problems for understanding the fruitfulness of the interdisciplinary work currently pursued in current neuroscience.

Associative bridge laws are perfectly consistent with the multiple-realizability of psychological kinds. Amygdala activation signals the engagement of processes involving negative emotions but, as discussed at length, it can also be triggered by other cognitive processes, such as the perception of sharp contours and unusual stimuli. In addition, processes involving negative emotions could be implemented in other neural locations. Still, as long as we can order these manifold inter-level interactions in a probabilistic way, and provided that we factor in the relevant task, neuroimaging data can be used in particular reverse arguments to discriminate among competing higher-order cognitive hypotheses. Similarly, pattern-based inferences are also compatible with multiple realizability, even in its most radical forms. In the example presented above, multi-voxel patterns can be used by classifiers to determine the engagement of recollection or familiarity-based recognition processes. The patterns are extracted and the classifiers are trained in specific tasks and for each subject individually. For instance, that some multi-voxel pattern is accurately categorized as a recollection process by

a classifier trained for a subject does not entail that the same pattern would be so categorized by a classifier trained on a different subject. Likewise, the fact that a classifier trained for a subject in a particular recollection/familiarity task is reliable, does not mean that it would still reliably distinguish between these processes in a different type of task—e.g., one that uses visual objects instead of words. In short, the successful use of these multi-voxel patterns and classifiers to discriminate between theories of the dynamics of recognition processes does not depend on whether they are stable across subjects or even, within certain limits, across tasks. Hence, the assumption that recollection and familiarity processes are multiply realizable leaves the applicability of context-sensitive reverse inferences completely unscathed.

5.3 Revisiting Marr-Levels and Reductionism

Let us conclude by discussing the third and most general implication of our account. The classic reductive model of interlevel relations and Marr's influential division of the study of cognition into three levels are, strictly speaking, independent. Early eliminative materialists such as Paul Churchland (1981) endorse reductionism while rejecting Marr-levels, whereas many philosophers recognize the usefulness of Marr-levels but eschew reductionism (Bechtel and Mundale 1999). However, the two views mutually support each other. To wit, a standard reductionist response to multiple realizability is to argue that antireductionists set up a straw man by selecting relata on the cognitive side that are too coarse-grained to be reduced (Kim 1992; Shapiro 2000).[14] The general idea underlying this response is that, as cognitive functions are progressively broken down into smaller subcomponents, it becomes more likely that we will reach a level where (local) reductive bridge laws can be established. Note how this picture of functional decompositions and local reductions fits in naturally with a standard interpretation of Marr-levels, according to which it only makes sense to ask about the lower-level implementation of functions once the cognitive processes that compute them have been laid out in algorithmic detail.

We do not deny that hypotheses regarding the neural implementation of cognitive-level processes constitute a significant portion of cognitive neuroscience. Indeed, astonishing progress has been made in the study of how certain perceptual and motor functions are carried out in the brain. However, we believe that this model of the psycho-neural interface as essentially addressing Marr-level 3 hypotheses is inadequate, as it leaves out much of the cognitive neuroscience of higher cognition. On the reductive account of Marr-levels, psychology and neuroscience only begin to meaningfully interact once we can ask how cognitive processes are implemented in neural hardware. This ignores a different—but

[14] For instance, a cognitive function such as 'language processing' is too coarse-grained to be directly associated to stable neural locations, as attempted by Poldrack (2006), to determine the reliability of inferences from activation in certain regions of Broca's area to the engagement of language processing. Still, the appropriate relata might be found, the reductionist insists, if we focus on subcomponents of language processes. For example, Pylkannen and colleagues (2011) have attempted to find, with some success, the neural correlates of certain semantic compositional operations, a key aspect of semantic processing.

equally important—type of psycho-neural interaction: using neural data to select among competing cognitive processes even when we have no clue how they could be neurally implemented (Del Pinal and Nathan 2013). This possibility of delving into the neural level only to 'come back up' to select hypotheses at the *cognitive* level is too often ignored by critics.

Our account of associative bridge laws also clarifies why, contrary to reductionist assumptions, it is often easier to employ neural data when Marr-level 2 hypotheses are not (yet) fully developed. For example, syntactic and semantic theories in linguistics are quite refined, but neuroimaging studies have been notoriously difficult to apply in this area. Linguists often face the task of determining whether a certain process is syntactic or semantic, with different models yielding different predictions. Take the case of 'it is raining,' used to mean that it is raining at the place of utterance. To account for this implicit location restriction, some models assume that a syntactic variable is inserted in the sentence prior to semantic interpretation (Stanley 2000); other models assume that the meaning of 'raining' is enriched to include the specification of a location (Recanati 2011). The former explanation appeals to a syntactic process; the latter to a semantic one. If we found bridge laws mapping syntactic and semantic operations onto distinct locations or patterns of neural activation, we could try to discriminate between the two models by scanning subjects while processing such sentences. Unfortunately, establishing the relevant bridge laws is proving to be a daunting task: since semantic and syntactic processes usually work in tandem, they are extremely hard to disentangle. As a consequence, we cannot, at present, use neural data to discriminate between syntactic and semantic models of ellipsis. In contrast, models of moral and economic decision making are still comparatively undeveloped. As Camerer and colleagues (2005) argue in great detail, one of the main divisions in current studies of decision making is between hypotheses that assume more rational processes, and hypotheses that assume an essential involvement of emotions. This division is illustrated by our discussion of moral decision making, and also emerges in several neuroeconomic debates, such as in competing explanations of the endowment effect (Knutson et al. 2008). This contrast is significant for the use of reverse inferences because we have bridge laws that map emotions and controlled rule-guided behavior onto distinct brain regions (Miller and Cohen 2001; Greene 2009). Consequently, we can often test these decision-making hypotheses using reverse inference. However, as this branch of science progresses and mixed models that incorporate both rational and emotional components become more common, it may become more difficult to use our current bridge-laws to discriminate amongst them in neuroimaging studies.

The occasional difficulty in finding bridge laws that discriminate between advanced Marr-level 2 models—compared to the relative ease with which such laws often discriminate more elementary models—is hard to reconcile with the traditional reductive interpretation of Marr's framework. Hypotheses that have an advanced functional decomposition are better suited for implementation; hence, from the reductive perspective, they should also be better candidates for interaction and integration with the neural level. Furthermore, since few of our

current hypotheses regarding capacities such as language or decision-making are ready for Marr-level 3 implementation, it is hardly surprising that those who accept the reductive interpretation of Marr levels typically endorse the relative autonomy of the psychology of higher cognition. In contrast, our dynamic account makes better sense of the current limitations and achievements of interdisciplinary research at the border of psychology and neuroscience. Once again, our approach is compatible with the possibility that science will eventually discover the neural implementation of higher-level cognitive processes. Yet, abandoning the reductive perspective suggests other significant ways in which neural data can be employed to advance psychology.

References

Anderson, M. (2010). Review of *Neuroeconomics: Decision Making and the Brain*, eds. Glimcher, Camerer, Fehr, and Poldrack. *Journal of Economic Psychology 31*, 151–54.

Bechtel, W. and J. Mundale (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science 66*, 175–207.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.

Bourget, D. and D. Chalmers (2013). What do philosophers believe? *Philosophical Studies*, 1–36.

Camerer, C. F., G. Loewenstein, and D. Prelec (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature 43*, 9–64.

Churchland, P. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy 78*(2), 67–90.

Coltheart, M. (2013). How can functional neuroimaging inform cognitive theories? *Perspectives on Psychological Science 8*(1), 98–103.

Del Pinal, G. and M. J. Nathan (2013). There and up again: On the uses and misuses of neuroimaging in psychology. *Cognitive Neuropsychology published online*(dx.doi.org/10.1080/02643294.2013.846254).

Enç, B. (1983). In defense of identity theory. *The Journal of Philosophy 80*, 279–298.

Fazekas, P. (2009). Reconsidering the role of bridge laws in inter-theoretic relations. *Erkenntnis 71*, 303–22.

Fodor, J. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese 28*, 97–115.

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Nous 31*, 149–63.

Fodor, J. A. (1999). Let your brain alone. *London Review of Books 21*.

Gallistel, C. R. (2009). The neural mechanisms that underlie decision making. In P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.), *Neuroeconomics: Decision Theory and the Brain*, pp. 419–24. Elsevier.

Gazzaniga, M. S. (Ed.) (2009). *The Cognitive Neurosciences* (Fourth ed.). Cambridge, MA: MIT Press.

Glimcher, P. W., C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.) (2009). *Neuroeconomics: Decision Making and the Brain* (1st ed.). London and Waltham, MA: Elsevier.

Glimcher, P. W. and E. Fehr (Eds.) (2014). *Neuroeconomics: Decision Making and the Brain*. Burlington, MA: Elsevier.

Greene, J. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed.)., Chapter 68, pp. 987–999. Cambridge, MA: MIT Press.

Greene, J., R. Sommerville, L. Nystrom, J. Darley, and J. Cohen (2001). An fMRI investigation of emotional engagement in moral judgment. *Science 293*, 2105–08.

Henson, R. (2005). What Can Functional Neuroimaging Tell the Experimental Psychologist? *Quarterly Journal of Experimental Psychology 58A*, 193–233.

Hooker, C. A. (1981). Towards a general theory of reduction. part iii: Cross-categorical reductions. *Dialogue 20*, 496–529.

Horst, S. (2007). *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*. New York: Oxford University Press.

Hutzler, F. (2013). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *Neuroimage in press*.

Kandel, E., J. Schwartz, T. Jessell, and S. Siegelbaum (2013). *Principles of Naural Science* (5th ed.). New York: McGraw-Hill.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research 52*, 1–26.

Kim, J. (1999). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2005). *Physicalism, Or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese 151*, 547–59.

Klein, C. (2009). Reduction without reductionism: A defense of nagel on connectability. *The Philosophical Quarterly 59*(234), 39–53.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics 4*, 143–62.

Knutson, B., E. G. Wimmer, S. Rick, N. G. Hollon, D. Prelec, and G. Loewenstein (2008). Neural antecedents and the endowment effect. *Neuron 58*, 814–22.

Legrenzi, P. and C. Umilta (2011). *Neuromania*. New York: Oxford University Press.

Lewis, D. K. (1969). Review of art, mind, and religion. *The Journal of Philosophy 66*, 23–35.

Lindquist, K. A., T. D. Wager, K. H., B.-M. E., and B. L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences 35*, 121–202.

Machery, E. (2013). In defense of reverse inference. *British Journal for the Philosophy of Science* (published online).

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Marras, A. (2002). Kim on reduction. *Erkenntnis 57*, 231–57.

Mather, M., J. T. Cacioppo, and N. Kanwisher (Eds.) (2013). *20 Years of fMRI—What Has It Done for Understanding Cognition*, Volume 8. *Perspectives on Psychological Science*.

Miller, E. K. and J. D. Cohen (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci. 24*, 167–202.

Miller, G. (2008). Growing pains for fMRI. *Science 320*, 1412–1414.

Nagel, E. (1961). *The Structure of Science*. New York: Harcourt Brace.

Norman, K., J. Quamme, and E. Newman (2009). Multivariate methods for tracking cognitive states. In K. Rosler, C. Ranganath, B. Roder, and R. Kluwe (Eds.), *Neuroimaging of Human Memory: Linking Cognitive Processes to Neural Systems*. Oxford University Press.

Norman, K., J. Quamme, and D. Weiss (2010). Listening for recollection: a multi-voxel pattern analysis of recognition memory retrieval strategies. *Frontiers in Human Neuroscience 4*, 1–12.

Phelps, E. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology 57*, 27–53.

Phelps, E. (2009). The study of emotion in neuroeconomics. In P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack (Eds.), *Neuroeconomics: Decision Making and the Brain*, Chapter 16, pp. 233–250. London: Academic Press.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences 10*(2), 59–63.

Poldrack, R. A. (2008). The role of fmri is cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology 18*, 223–27.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inferences to large-scale decoding. *Neuron 72*(692-97).

Putnam, H. (1967). Psychological predicates. In W. Capitan and D. Merrill (Eds.), *Art, Mind, and Religion*, pp. 37–48. Pittsburgh, PA: University of Pittsburgh Press.

Pylkkanen, L., J. Brennan, and D. Bemis (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Language and Cognitive Processes 26*(9), 1317–37.

Recanati, F. (2011). *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Ross, D. (2008). Two styles of neuroeconomics. *Economics and Philosophy 24*, 473–83.

Satel, S. and S. Lilienfeld (2013). *Brainwashed: The Seductive Appeal of Mindless Neuroscience*. New York: Basic Books.

Shapiro, L. A. (2000). Multiple realizations. *The Journal of Philosophy 97*(12), 635–54.

Stanley, J. (2000). Context and logical form. *Linguistics and Philosophy 23*, 391–434.

Uttal, W. R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes*. Cambridge, MA: MIT Press.

Uttal, W. R. (2002). Precis of the new phrenology: The limits of localizing cognitive processes in the brain. *Brain and Mind 3*(2), 221–28.

Van Orden, G. C. and K. R. Paap (1997). Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain. *Philosophy of Science 64*, S85–94.

To be presented at the PSA meeting in Chicago, November 2014

Physicalism, Introspection, and Psychophysics: The Carnap/Duncker Exchange

Uljana Feest
Leibniz Universität Hannover
October 2014

In 1932, Rudolf Carnap published his article "Psychology in a Physical Language." The article prompted a critical response by the Gestalt psychologist Karl Duncker. The exchange is marked by mutual lack of comprehension. In this paper I will provide a contextualized explication of the exchange. I will show that Carnap's physicalism was deeply rooted in the psychophysical tradition that also informed Gestalt psychological research. By failing to acknowledge this, Carnap missed out on the possibility to enter into a serious debate and to forge an alliance with a like-minded psychologist at the time. I conclude by suggesting that the kind of physicalism practiced by Gestalt psychologists deserves to be taken seriously by current philosophy of psychology.

In the early 1930s, Rudolf Carnap laid down his project of overcoming metaphysics by means of linguistic analysis (Carnap 1931a) and specified a universal (physical) language as the language of choice (Carnap 1931b). It is well known that Carnap's 1931b article gave rise to what is often referred to as the "protocol-language debate" within the Vienna Circle (Neurath 1932; Carnap 1932b). While there is some impressive historical and philosophical scholarship about this debate (e.g., Uebel 2007), one strand of it has not received much attention, namely the ways in which Carnap's views about the physicalizability of protocol sentences were related to research in experimental psychology at the time. This is especially surprising given the fact that Carnap, in his article "Psychology in a Physical Language" (1932a) attempted to spell out the implications of this view for psychology. This article was met by a critical response by the psychologist Karl Duncker (1932), which in turn prompted a reply from Carnap (1932c). The exchange is characterized by a surprising degree of mutual incomprehension, with Duncker suggesting that Carnap's critique of (introspective) psychology was attacking a strawman and Carnap saying that Duncker had completely missed his point.

In this article I will explicate and contextualize the exchange between Carnap and Duncker. I will shed some light on the reasons why the two talked past each other and I will show that Duncker did put his finger on the fact that (1) Carnap's position failed to address scientific practice, and

To be presented at the PSA meeting in Chicago, November 2014

that (2) Carnap did indeed attack several strawmen. I will lay out that Carnap's turn to a physical language was motivated by his aim to provide an objective foundation for protocol sentences (section 2), and argue that the way in which Carnap executed his project of physicalizing protocol-sentences was deeply informed by psychophysics (section 3). I will then (in section 4) turn to Carnap's 1932a article "Psychology in a Physical Language," where he claims to be addressing the implications of his views for psychology. Section 5 details Duncker's response and Carnap's answer. Finally, in section 6, I will draw out two underlying issues in this debate, i.e., (a) the status of introspection in psychological research, and (b) the question what (if any) metaphysical presuppositions were made by psychophysical research.

2. Overcoming Metaphysics and the Problem of Experience

In his famous article "Overcoming Metaphysics" Carnap laid out the very lean conception of philosophy characteristic of the Vienna Circle (Carnap 1931a). According to it, philosophy was neither to engage in metaphysical speculations about age-old topics, nor in naturalistic treatments of them. Rather, it was essentially reduced to providing meta-analyses of existing discourses in order to clean them of "pseudo-sentences;" sentences that look grammatically like sentences, but are in fact meaningless. The method of choice (logical analysis of language) was to proceed by translating every sentence that is formulated in the so-called "material mode of speech" into a sentence in the "formal mode of speech" (a sentence about a sentence). This method was to reveal whether a given statement was logically consistent and empirically meaningful.

In response to the question of what it takes for a statement to be empirically meaningful, Carnap introduced a version of the well-known verificationist criterion of meaning that we still associate closely with the doctrine of logical positivism. According to it a word ***a*** is meaningful only if (1) empirical indicators for a are known, (2) it is known what protocol sentences the sentence S(***a***) can be derived from, and (3) the path towards verifying S(***a***) is known (Carnap 1931a, 224). Carnap's verificationist semantics for words emphasized the empirical truth conditions for sentences in which the words occur. These empirical truth conditions were provided by "observational" or "protocol-sentences" and he stated that while there was to date no agreement

about the form or content of such sentences, they were commonly thought to refer to something that is "given" (Carnap 1931a, 222).

This raised the question of what were criteria of meaningfulness of protocol sentences themselves? Was their meaningfulness ensured by a primitive notion of the "given," i.e., by the subjective experience that is – in the material mode – reported by protocol sentences? Or was there a more 'public' way of stating the truth conditions for protocol sentences? It is this question that Carnap addressed in his "Physical Language as the Universal Language of Science" (Carnap 1931b), where he argued that both protocol sentences and "system sentences" (i.e., sentences capable of being derived from, and verified, by protocol sentences) are part of an overarching language: the universal language of science. By the requirement of *universality*, Carnap meant that such a language "can describe every state of affairs" (Carnap 1963 [1932a] 400). But in addition he argued that such a language should also be *intersubjective,* i.e., it should be usable by everybody in the same way. It is in this second respect that Carnap's aim in this work differed from the one proposed in his 1928 *Aufbau*, where he had wanted the universal language to be that of subjective experience. It was because of the requirement of *intersubjectivity* that Carnap turned to the physical language as the universal language (see Uebel 2007).[1]

Carnap's thesis that (seemingly subjective) protocol sentences were translatable into the (intersubjective) language of physics was ostensibly part of a larger thesis, namely that *all* sentences are translatable into sentences of the physical language. Surprisingly, Carnap did not present an argument for this larger thesis, except to suggest that its truth was obvious, at least in the case of "the inorganic sciences" (chemistry, geology, astronomy) and even biology since they were (in the material mode) dealing with physical objects. However, since his main target was the physicalizability of protocol sentences, a separate argument was required, and he attempted to provide one in sections 4 of his "Physical Language."

[1] In his "Physical Language" article he still maintained that protocol-sentences were the most basic sentences of science that could not themselves be doubted (438), but in response to Neurath's critique, he revised this position to say that any scientific sentence within a physicalist system could function as a protocol sentence ("Über Protokollsätze" 224) and no sentence could function as an ultimate epistemic basis (225).

To be presented at the PSA meeting in Chicago, November 2014

3. The Psychophysical Underpinnings of Physicalized Protocol Sentences

Carnap's argument for the translatability of protocol sentences into physical sentences took the form of an empirical claim: He posited that it is possible to find a quantitative equivalent for every qualitative (protocol) sentence, and he noted that this is not a logical necessity but simply a contingent empirical fact, such that there is a functional dependency between protocol sentences and physical sentences. (In the material mode, it is an empirical fact about the way in which our experience is structured in relation to the physical world). From this, Carnap derived the claim that it is possible to find a corresponding physical sentence for every protocol sentence, a fact that provides the basis of the very possibility of practicing an intersubjective physics.

It is clear that Carnap expected the physical sentences in question to be sentences about the brain, but realized that neuroscience at the time did not deliver sentences that directly corresponded to sentences about experiences. For this reason, he decided to describe the relevant brain states in terms of behavioural dispositions, specifically, the disposition to display particular behaviours in response to particular stimuli. This was made explicit in his subsequent "Psychology in a Physical Language," but is already apparent in the way he explains his position in his 1932b paper. For example, he states that it is possible to establish empirical correlations between the qualitative statements of protocol sentences and the quantitative determination (in a physical language) of the conditions under which they are uttered. For example, when examining color vision, he argued, one has to vary "the physical conditions (e.g., the combination of various frequencies of oscillations) and discover the conditions to which S reacts with the protocol statements containing the qualitative terms in question" (Carnap 1963 [1932]; 408). Then he states "The discovery of the set of these physical terms corresponding to a definite qualitative term will be called the 'physicalizing" … of this qualitative term" (ibid.).

Unfortunately, Carnap does not provide a reference for this, but it is striking that there was in fact a research program that did just what Carnap was describing: i.e., vary physical stimulus conditions and measure responses. This research program, psychophysics, had famously been founded by Fechner (1860) and continues to be active until today (Heidelberger 2004). At the time of Carnap's work in the 1930s, famous proponents of this research were members of the Berlin/Frankfurt school of Gestalt psychology, with whom Carnap was at least indirectly

acquainted (Feest 2007).[2] I therefore suggest that Carnap was aware of psychophysical research and that his proposal to translate protocol sentences into sentences about behavioural dispositions was in fact inspired by this tradition.[3] With this claim I do not wish to refute the common assumption that the position advocated here Carnap is a kind of logical or analytical behaviourism. My thesis, rather, is that Carnap's articulation of this form of behaviourism (i.e., what Carnap called the "physicalizing" of qualitative terms), relied on a contingent empirical fact. And the way in which he imagined the empirical investigation of this fact was practiced by a particular (at the time quite dominant) approach within psychology. The implication I want to highlight here is that specific attempts to translate a psychological sentence (Mr A is angry) into physical sentence (Mr. A exhibits particular behaviors) are going to build on psychophysical research, which in turn will necessarily involve first-person data.

4. Physical Language, Physicalized Observation Sentences and Psychology

In his article "Psychology in a Physical Language" Carnap explored the implications of his physicalism for psychology as a science, specifically focusing on the question of the kinds of observation sentences were admissible in psychology. His targets were "observations of the mental states of others" (section 3) and "self-observations" (section 7).

With regard to observation-statements about the mental states of others (e.g., "Mr A is angry"), Carnap argued that insofar as such sentences are meaningful at all, they are only meaningful if they can be translated into statements about physical behaviours (about Mr. A's disposition to behave in certain ways). This allows for the derivation of sentences that state truth conditions for the sentence in question (about Mr. A's actual behaviour), though (as Carnap lays out) to treat them as such requires an additional premise, namely that in general people display the behaviors in question when angry. Carnap uses this to argue that if we want to test a particular sentence about the content of someone's mind, we have to (a) appeal to a general sentence about the kinds of physical conditions that need to be in place when we use the term "anger" and (b) point to a

[2] When discussing (both in his 1928 *Aufbau* and in his 1932 article about physical language) the question what are the units of experience that protocol language typically describes, he opted for those identified by Gestalt psychology.

[3] Even if this historical thesis does not hold up, I maintain that psychologists at the time would have recognized the similarity (and that's all that matters for my subsequent argument).

particular set of physical conditions as instantiating the general conditions in question. Carnap then contrasts this ("rational") mode of justification with one where the emotional state of Mr. A is ascertained in a more "intuitive" way. He argues, however, that such intuitive sentences are either meaningless or can be translated into one that states the physical conditions that provide it with meaning. In section 7, he picks up on this claim and says that the same is true of sentences about our own mental states: For example, when we utter a sentence like "I am nervous right now," this sentence is either meaningless or its meaning is provided by empirical truth conditions in a physical language (shaky hands, sweaty palms, etc.).

One might wonder whether (and if so, how) these considerations were relevant to the research practices of experimental psychology, as Carnap seems to suggest. In a nutshell, he had two answers to this: First, he claimed that by failing to appreciate his point about the semantics of psychological sentences, psychologists were prone to falling into a kind of psychophysical dualism (which is apparent, for example, when I say that I am in physical state X and *in addition* in mental state Y). Second, he cautioned against attributing a special kind of epistemic significance to first-person experiences (of other minds or of one's own mind), pointing out instead that in science one always deals with *sentences* about experiences, which should be treated as the behavioural outputs of complicated detection devices under particular physical conditions: "In principle, there is at most a gradual epistemic difference between the utterances of a fellow human being and a barometer" (p.124; translation by me). (see also p. 140 for similar statements).

Carnap repeatedly comments on the confused state of the then current psychology (for example "understanding" and "introspective" psychology), but he never quite says who is actually committing the above two errors. It is possible that he had in mind Wilhelm Dilthey, who had argued for understanding as a first-person mode of access to the subject matter of the human sciences. But Carnap's critique is a little confusing by virtue of the fact that Carnap also throws in a different psychological approach (again without mentioning any names), which studies "purposeful behaviour." I suggest that here Carnap had in mind the American neo-behaviorist Edward Chace Tolman, whom Carnap probably met at Vienna Circle meetings (or at least knew about via Egon Brunswick).[4] Tolman emphasized the goal-directed nature of much behaviour and

[4] Tolman's book, *Purposive Behavior in Animals and Men* also appeared in 1932.

essentially introduced proto-cognitive mental states to explain them. With respect to this explanatory practice, Carnap argues that it is acceptable as long as we realize that talk about purposes can be fully physicalized, i.e., translated into a language that specifies lawful regularities between specific stimuli and behavioural dispositions. It bears stressing that Dilthey's approach was quite different from Tolman's, insofar as the former talked about a mode of empirical access (understanding), whereas the latter talked about an explanatory concept (purpose). Hence, we can note that Carnap's critique of psychology was fairly broad.

While this does not come out in Carnap's article, one other likely target of Carnap's attack on psychology was the psychologist Karl Bühler, who was based in Vienna at the time. In 1927, just a few years before Carnap's "Psychology in a Physical Language", Bühler had published a book about what he had termed the "crisis of psychology" (Bühler 1927). In this book, Bühler specifically attacked the physicalism of Gestalt psychologists like Wertheimer, Köhler, and Koffka. According to Christoph Limbeck (2014), Carnap presented his ideas about a physicalist psychology in Bühler's colloquium on two separate occasions in the summer of 1930, where they gave rise to heated discussions.[5] In the light of the hypothesis that Carnap's critique of contemporary psychology may have been directed at Bühler's 1927 book, and the light of the fact that Bühler, in this book, specifically opposed what he took to be the "physicalism" of the Berlin school of Gestalt psychology, it is not surprising that members of this school took an interest in Carnap's position.

5. The Carnap/Duncker Exchange

Carnap's 1932a article prompted a reply by Carl Duncker, a younger member of the Berlin school of Gestalt psychologists. This reply (and Carnap's subsequent response) demonstrates a deep mutual incomprehension. This is especially surprising in the light of my above thesis that Carnap's physicalism was informed by the psychophysical tradition, and that Gestalt psychological research has to be placed in that tradition. In this section, I provide a brief overview of the exchange, followed (in section 6) by an elaboration of my thesis.

[5] I would like to thank Christoph Limbeck for drawing my attention to this!

Even though it is clear, especially at the beginning of his article, that Duncker had misunderstood some of Carnap's points, he did object to Carnap's characterization of the two supposed problems of psychology, namely the danger of falling into a mind-body dualism and the tendency to attribute too much epistemic significance to introspective data. Carnap had argued that these two problems could be avoided if one took the general stance of behaviorism: "The position advocated here is essentially in agreement with the psychology known as 'behaviorism'" (Carnap 1932a, 124, translation by me).[6] Duncker argued that Gestalt psychological findings and methodology were more congenial to what Carnap was aiming at.

As mentioned above, Carnap claimed that his analysis of psychological sentences as translatable into physical ones could help psychologist avoid the inference that the two types of sentences referred to two separate kinds of "things." Duncker responded with utter incomprehension, stating that he was not aware of many contemporary psychologists who made substantial metaphysical assumptions about mind and body (Duncker 1932, 165), and that Gestalt psychologists were in fact physicalists, though he did not think that this had any implications for the goals or methods of non-behaviorist psychology (ibid. 176). He backed this up by claiming that psychophysics also viewed organisms as detection devices, like Canap, and that they endorse a principle of isomorphism between mental and physical processes (Duncker 1932, 174). However, Duncker also used some careless formulations, which seemingly contradicted his anti-metaphysical stance. For example, he asserted that it was possible to conceptualize anger both as a behavioural disposition and as an inner experience. For Carnap these assertions showed Duncker to be falling in exactly the metaphysical trap that Carnap had warned about (Carnap 1932b, 186/7). Accordingly, Carnap responded with some surprise about Dunker's claim to be a physicalist, noting that he and Duncker clearly had in mind very different notions of *physicalism*: Carnap's physicalims was about the translatability of psychological language into physical language, whereas Duncker's physicalism was about finding the physical basis of introspectively accessible experience.

This brings us to Carnap's second critique of psychology, namely that of introspective methods. Carnap's rejection of introspection was closely related to his above-mentioned point about the dangers of being misled into a dualist metaphysical position. It was not aimed at the usage of this

[6] Here Carnap mentions that J. B. Watson's *Der Behaviorismus* had been translated into German in 1930.

method as such, but at certain interpretations of its results, with its inherent danger of reifying the object of introspection, a point he reiterates in his reply to Duncker. In his response, Duncker stated that (a) Gestalt psychology (like all of psychophysics) relied on some kinds of self-reports since it was not clear how psychophysical laws could be formulated otherwise, but (b) that one did not have to be a behaviorist to reject the assumption that psychology aims at the "absolute content of a quale" (Duncker 1932, 167). Moreover, (c) when psychophysicists investigated (for example) color vision, they were not interested in the subjective experience of the color as such, but in their "order properties" ("Ordnungseigenschaften"). For example, if an individual experienced inverted qualia, this was irrelevant for the language of science, so long as that person responded to all the same stimuli in the same way as a person with 'normal' experiences. All of these points then lead up to Dunckers's somewhat exasperated question: "I ask once again, which 'interpretation' of introspection is Carnap arguing against?"(Duncker 1932, 169).

Summing up, Duncker held that an endorsement of introspection was compatible with a lean metaphysics and with the notion of biological organisms as physical detection devices. Yet, at the same time, he held that introspection was an indispensable tool for the empirical investigation of such devices. Carnap, in his reply, repeated that he did not care what tools psychologists used, so long as the resulting sentences could be stated in a physical language. Thus, he took Duncker's conjectures to be irrelevant to his point.

6. Some Underlying Issues

In this section, I will argue that while it is true that Duncker misunderstood Carnap's point, it is also the case that Carnap was not receptive to the implications of Duncker's remarks, namely (a) that Carnap's physicalism did not have much practical relevance to at least some non-behaviorist psychology as it was being practiced at the time (or to the extent that it did, Carnap had failed to engage with the targets of his critique), and (b) that Carnap's physicalism was rooted in psychological research and hence he was well advised to engage with the foundations of that research.

6.1 The Heuristic Function of Introspection

Clearly, one issue at play in the exchange between Darnap and Duncker was a version of the distinction between discovery and justification, with Carnap declaring scientific methodology irrelevant to the epistemological status of psychological claims such as "Mr A is angry." At that point in time, questions surrounding the production and epistemic status of first-person experiential reports had been debated within experimental psychology for well over half a century, following the work of Fechner, Brentano, G.E. Müller, and many others. The fact that Carnap simply ignored these debates must have seemed bizarre to a practicing scientist like Karl Duncker. Now, Carnap might have replied that those very discussions were themselves rooted in metaphysical assumptions, and that it was precisely for this reason that he was aiming at a purely formal analysis of scientific language rather than the metaphysical presuppositions of scientific practice. But the question is whether Carnap was entitled to this stance, since – as I argued above – his very conception of the physicalization of protocol sentence took for granted the project of psychophysics, that is, the project of discovering functional relationships between physicalist and mentalist descriptions. It is this fact that Duncker is also alluding to when arguing that even though experimental psychology mainly varies physical stimuli, this activity is heuristically guided by introspection (Duncker 173). This can be illustrated if we go back to Carnap's claim (see section 4 above) that the empirical conditions of application for a psychological sentence (such as "Mr. A is angry") were provided not only by an individual displaying the relevant behaviors, but also by an empirical law that described the types of behaviors typically displayed by angry people. Such laws, Carnap tells us, are the result of inductive generalizations. What Duncker is pointing out (on my construal) is that such generalizations are based on self-reports, and that therefore human subjectivity cannot be eliminated from the research process.

Again, Carnap might have replied that it was precisely the (merely) heuristic nature of introspection that rendered it irrelevant to serious logical analysis, and moreover, that even in the actual research (as he and Duncker agreed) only the introspective *reports* (not the introspective experiences themselves) counted. But even if we grant this, I would like to suggest that some of the unproductive harshness of this exchange could have been avoided if Carnap had acknowledged his debts to the psychophysical tradition. This might have helped him understand why Duncker was so irritated by Carnap's positive assessment of methodological behaviorism, since after all behaviorism, by virtue of not talking about mental states at all (physicalized or not)

To be presented at the PSA meeting in Chicago, November 2014

radically rejected the very method that Carnap's physicalism was built on, namely that of treating experiential reports as relevant scientific data. I will now argue that it might also have helped him provide a more nuanced description of the kinds of mind-body parallelisms available at the time.

6.2 Varieties of Mind-Body Parallelisms

As indicated above, Duncker rejected Carnap's claim that contemporary psychology's use of the material mode was leading it down the road to mind-brain dualism. But what did Duncker have in mind here?

Carnap's turn to the analysis of language was motivated by his aim to avoid metaphysics. While this specific program of antimetaphysics is particularly well known, the mid- to late 19th century had seen a lot of debates about banning metaphysics from scientific and philosophical discourses. In this vein, already Fechner, in his 1860 *Elements of Psychophysics*, had formulated an account of the psychophysical relationship that aimed to steer clear of fruitless debates between materialists and idealists at the time. His response was to argue that mental and physical properties were token identical, but depending on one's perspective, one could only ever empirically apprehend one or the other and never both at the same time (Fechner 1860; Heidelberger 2004ab). While this type of position was often referred to as a kind of "parallelism" (see also Heidelberger 2003), a more apt description might be "dual-aspect theory," since this term does not suggest the existence of distinct substances or properties, but merely of distinct perspectives. It was precisely this notion that underwrote Fechner's empirical project of psychophysics. As Heidelberger (2003) explains, it is possible to distinguish between three layers of Fechner's parallelism: an empirical hypothesis about the functional relationship between physical and psychological descriptions, a dual-aspect theory of the mind/body relationship, and a cosmological thesis, according to which even inorganic processes have a mental side to them. I have argued above that Carnap not only shared Fechner's empirical hypothesis, but also his methodology of how to investigate this functional relationship (by varying stimuli and recording responses). While he clearly did not agree with Fechner's mind-body theory (let alone with his cosmological thesis), it bears stressing that Fechner's mind-body theory was not a kind of substance dualism. Rather, it was a dual aspect theory which scholars like Duncker may well

have taken to lay the foundations for the very possibility of the psychophysical research that Carnap implicitly appealed to, when casting protocol sentences in terms of behavioural dispositions.

Now, it is clear that for Carnap, the linguistic description of behavioural dispositions (expressed in a physical language) merely ensured the meaningfulness of psychological sentences, whereas for Fechner they expressed psychophysical laws, i.e., laws that describe the empirical relationship between two types of the magnitudes: experiences and physical stimuli. But given that Carnap's semantic analysis also exploited (or at least assumed) the empirical relationship in question, it is well worth pointing out that Fechnerian psychophysics attempted to account for the existence of the empirical relationship without positing separate substances or even properties. By stating that the correspondence between the two languages was a crude empirical fact, Carnap may have been able to avoid metaphysical speculation, but there is also a sense in which this appeal is somewhat unsatisfactory. Moreover, it remains to me an open question to what extent Carnap's thinking about this was implicitly informed by some version of the dual-aspect theory that had underwritten Fechner's psychophysical research. Whether or not this was the case, I argue that for Duncker this may have been an intuitive way to think about the matter, which would account for his difficulties in comprehending Carnap's point.

7. Conclusion

I have argued that Carnap's account of protocol sentences (including those of psychology) was deeply informed by the psychophysical research tradition of the mid-19th to early 20th centuries. In the light of this, I have provided an analysis of the exchange between Carnap and Duncker, in which Duncker questioned Carnap's contention that the methodological approach of behaviorism within psychology was congenial to his approach, arguing instead that Gestalt psychology came much closer to Carnap's outlook. I substantiated Duncker's assessment by providing a reading of Duncker's analysis that highlights the following two points: First, Gestalt-psychological research (and psychophysical research more generally), while giving a lot of weight to first-person experiential reports, did not necessarily invest them with epistemic certainty or treat them as being about irreducible qualia. Second, researchers in the psychophysical tradition (including

Gestalt psychologists) were not necessarily committed to a mind-brain dualism, even if they aimed to formulate psycho-physical laws.

Given that Carnap's formal analysis relied on (or at least presupposed the possibility of) this research, I argue that it was unwise for Carnap to reject as irrelevant Duncker's points, both because it unnecessarily alienated a potential psychological ally and because it prevented Carnap from acknowledging the extent to which he and Duncker shared similar philosophical roots. It also prevented him from recognizing that his project of physicalizing protocol sentences (in the formal mode) relied on research that granted some epistemic authority to subjective experience (in the material mode).

In conclusion I argue that the way in which Carnap tried to insulate his philosophical project, as concerned with the "epistemological status" of psychological sentences, from the question of how such sentences are established was part of a general trend away from being concerned with scientific practice. While philosophy of psychology in the 19th century had still been fairly practice-oriented (as evidenced, for example, in Fechner's 1860 *Elements of Psychophysics* or Brentano's 1874 *Psychology from an Empirical Standpoint*), the philosophical turn to formal analysis (along with the rise of behaviorism in the US), for some decades eclipsed the fact that much of psychology continued to make some use of first-person reports. As a result philosophers of psychology have only recently started to turn their renewed attention to questions about the meaning, role and justification of first-person reports in psychology's research practices. In this context, the physicalist analyses of introspection, as they were provided by advocates of Gestalt psychology, are still well worth considering. (Feest 2014)

REFERENCES

Bühler, K. (1927): *Die Krise der Psychologie*. Jena: Verlag Gustav Fischer.

Carnap. R. (1931a): Überwindung der Metaphysik durch logische Analyse der Sprache. *Erkenntnis* 2, 219-241

------ (1931b): Die physikalische Sprache als Universalsprache der Wissenschaft. *Erkenntnis* 2, 432-465

------ (1932a): Psychologie in physikalischer Sprache. *Erkenntnis* 3, 107-142

------ (1932b): Erwiderung auf die vorstehenden Aufsätze von E. Zilsel und K. Duncker. *Erkenntnis* 3, 177-188

----- (1963): The Physical Language as the Universal Language of Science. W. Alston & G. Nakhnikian (Ed.): *Readings in Twentieth-Century Philosophy*. New York: The Free Press, pp. 393-424. (originally published as Carnap 1931b)

Duncker, K. (1932): Behaviorismus und Gestaltpsychologie (Kritische Bemerkungen zu Carnaps ‚Psychologie in physikalischer Sprache'). *Erkenntnis* 3, 162-176

Feest, U. (2007): Science and Experience/Science of Experience: Gestalt Psychology and the Anti-Metaphysical Project of the Aufbau. *Perspectives on Science* 15(1), 38-62.

Feest, U. (2014): Phenomenal Experiences, First-Person Methods, and the Artificiality of Experimental Data. *Philosophy of Science* (in press)

Heidelberger, M. (2003): The Mind-Body Problem in the Origin of Logical Empiricism: Feigl and Psychophysical Parallelism. P. Parrini, M. Salmon, W. Salmon *Logical Empiricism: Historical and Contemporary Perspectives*, Pittsburgh, PA: Pittsburgh University Press

------. (2004a): Fechner's (Wider) Conception of Psychophysics – Then and Now. Contribution to "Fechner Day 2004" XXth Meeting of the *International Society for Psychophysics.* Coimbra, Portugal, 18-22 October, 2004 (unpublished)

------. (2004b) *Nature From Within: Gustav Theodor Fechner and His Psychophysical Worldview*. Pittsburgh: University of Pittsburgh Press. (Translated from the German by C. Klohr)

Limbeck, Christoph (2014): Der Physikalismus bei Bühler und Carnap. (unpublished manuscript).

Neurath, O. (1932): Protokollsätze. *Erkenntnis* 3, 204-214

Uebel, T. (2007): Empiricism at the Crossroads. The Vienna Circle's Protocol-Sentence Debate. Chicago: Open Court.

CETERIS PARIBUS LAWS AND *MINUTIS RECTIS* LAWS

LUKE FENTON-GLYNN
DEPARTMENT OF PHILOSOPHY, UNIVERSITY COLLEGE LONDON
GOWER STREET, LONDON, WC1E 6BT, U.K.

Abstract. Special science generalizations admit of exceptions. Among the class of non-exceptionless special science generalizations, I distinguish (what I will call) *minutis rectis* (*mr*) generalizations from the more familiar category of *ceteris paribus* (*cp*) generalizations. I argue that the challenges involved in showing that *mr* generalizations can play the law role are underappreciated, and quite different from those involved in showing that *cp* generalizations can do so. I outline some potential strategies for meeting the challenges posed by *mr* generalizations.

1. Introduction

Many philosophers of science speak as though all non-exceptionless scientific generalizations that (appear to) play at least some aspects of the law role (counterfactual support, inductive confirmation, predictive/explanatory import) tolerably well can be classed as *ceteris paribus (cp) laws*. The following are representative quotations:

> "A nonstrict law is a generalization that contains a *ceteris paribus* qualifier that specifies that the law holds under 'normal or ideal conditions,' [...]. The generalizations one finds in the special sciences are mostly of this kind. In contrast, a strict law is one that contains no *ceteris paribus* qualifiers; it is exceptionless not just *de facto* but as a matter of law." (Lepore and Loewer 1987, 632)

> "cp lawfulness is just a species of nomological necessity, the other species of nomological necessity being strict lawfulness. What distinguishes the two

> species is just that cp laws can have ... exceptions and strict laws can't" (Fodor 1991, 31–32)

> "Special science laws ... are usually taken to 'have exceptions', to be 'non-universal' or 'to be *ceteris paribus* laws'." (Reutlinger et al. 2011)

In each of these quotations, the notion of a non-exceptionless law is run together with that of a cp law. It is easy to find further confirmation that this running-together is widespread (see, e.g., Schurz 2002, 351, Schrenk 2007, 221, Woodward 2002, 303–304).

The identification of non-exceptionless 'laws' with cp laws can hardly be a matter of stipulative definition. The notion of a cp law has a richness that significantly outstrips the bare notion of *a law that admits of exceptions*. For one thing, the notion of a cp law is associated with a distinctive account of how exceptions arise.

A cp law is supposed to be endowed with an implicit or explicit clause that specifies that it holds 'other things being equal', where this latter notion is usually parsed in terms of the obtaining of 'normal' or even 'ideal' conditions (see Cartwright 1983, 46, Schurz 2002), and explicated in terms of the absence of significant difference-making interference from outside the system that the law in question seeks to characterize (see, e.g., Fodor 1989, 69n, Schurz 2002, 366-370). Exceptions are taken to arise due to the non-satisfaction of this cp clause.[1]

After reviewing the notion of a cp law (Section 2), I will argue (Section 3) that it is a mistake to equate non-exceptionless laws with cp laws: there is a distinct type of non-exceptionless law – which I will call a *minutis rectis (mr) law* – which admits of exceptions that aren't explained by the non-satisfaction of a cp clause. I will argue (Section 4) that mr laws pose a distinctive set of philosophical challenges. Finally (Section 5), I will examine some potential responses to these challenges.

[1]I'm skating over some differences between the various characterizations of cp laws that appear in the literature (for an overview, see Reutlinger et al. 2011). Some (e.g. Schurz 2002) distinguish different *types* of cp law. For present purposes, it suffices to note that the notion of a *minutis rectis* law that I will distinguish below is not a type of (or variant on the notion of a) cp law.

A terminological point: talk of cp (and mr) 'laws' is rather awkward in the context of a discussion of whether, and to what extent, the exception-ridden generalizations of the special sciences play various aspects of the law-role. I'm sympathetic to the objections that some philosophers (e.g. Woodward 2005, Woodward and Hitchcock 2003a,b) have to such law-speak. Nevertheless, because law-speak is so common in the literature, I shall not try to forgo it in what follows, and I shall drop the jarring scare-quotes when I use it.

2. Ceteris Paribus Laws

In ecology, one standard equation used for predicting population growth is the Logistic Equation (LE):

(LE) $$\frac{dn}{dt} = r_c n\left(1 - \frac{n}{K}\right)$$

Here n is the number of individuals in the population, $\frac{dn}{dt}$ is the growth rate of the population (the change in n, with respect to time t), r_c is the *intrinsic per capita growth rate* of the population (the growth rate that obtains in the absence of intra-species competition for resources), K is the *carrying capacity* (the maximum sustainable population size).

LE implies that when the population n of a species in a particular habitat is very small (so that there is little intra-specific competition for resources), the actual population growth rate $\frac{dn}{dt}$ is close to the intrinsic per capita growth rate r_c multiplied by the number n of individuals. But, as the population grows, the actual growth rate declines linearly (due to increasing competition). This decline continues until the carrying capacity K is reached, at which point population growth is 0.

It is an open question whether ecological generalizations – such as LE – should be called 'laws' at all. But they do appear to play certain aspects of the law role to at least some degree. Ecologists apply LE to certain populations (especially populations that aren't subject

to significant *inter*-species competition or predation) in order to make predictions, and to give explanations.[2]

LE holds only *ceteris paribus* because there are possible background conditions under which it is violated (even when applied to populations concerning which, in normal circumstances, it is predictively accurate). For example, it will not hold in the event of the population being subject to a cull, or in the event of a natural disaster that destroys (a large part of) the population. While LE may give accurate predictions about population growth *after* some such events, it won't accurately predict growth *during* such episodes. It simply doesn't include variables that represent such events.

Culls, etc., produce circumstances in which other things are *not* equal: interfering factors are present, so LE doesn't even approximately hold. An ecologist presumably wouldn't seek to model such factors, since they are not *ecological* factors. They interfere with the sorts of system that the ecologist seeks to model (viz. *ecosystems*), but come from 'outside' such systems. Perhaps this means that ecological generalizations will in-principle remain cp generalizations (compare Davidson 1970, 94, Fodor 1989, 69n).

There have been several attempts (e.g. Lepore and Loewer 1987, Fodor 1989, 1991, Woodward and Hitchcock 2003a,b, Woodward 2005) to show that generalizations like LE, despite holding only cp, can support counterfactuals, and sustain predictions and causal-explanatory relationships and thus play the law role to a non-negligible degree.

[2]Tsoularis and Wallace (2002) survey some successful applications of LE in ecology. Ecologists sometimes appeal to more complex equations than LE. The following discussion also applies to these more complex equations. In general, ecologists have an armory of equations (or systems of equations – i.e. models) for predicting population growth and other phenomena. Different models are more or less predictively successful with respect to different populations. The fact that such equations (or models) apply only to some populations – and even then only approximately – may disincline you to call them 'laws'. I'm sympathetic. (Though note that there is a nuanced literature in (philosophy of) ecology about whether there are genuine ecological laws: see, e.g., Colyvan and Ginzburg 2003, Lawton 1999, Turchin 2001.) And since this story is repeated throughout the special sciences, you may be disinclined to admit the existence of special science laws at all (except, perhaps, in a few special cases). Again, I'm sympathetic. To reiterate: the question with which I'm concerned is not whether such generalizations deserve to be called 'laws', but whether and to what extent they are able to do things like predict, explain, and support counterfactuals. As we'll see, LE is the sort of thing that Woodward and Hitchcock (2003a,b) and Woodward (2005) call an 'invariant generalization'. I have no objection to their alternative terminology.

Woodward and Hitchcock (2003a,b) argue that generalizations like LE support causal-explanatory relations because they are *invariant under a range of hypothetical interventions*.[3] For example, if we were to intervene upon the intrinsic growth rate r_c of the population (e.g. by genetic engineering to increase fertility), upon the carrying capacity K (by improving or depleting the environment), or upon the population size n (by carrying out a cull), then the actual growth rate, $\frac{dn}{dt}$, – *after the intervention episode* – would accord with LE.

The reason that LE 'supports' these interventionist counterfactuals is that, in evaluating them, we are considering the 'closest worlds' in which such interventions occur (see Hitchcock 2001, 283; compare Lewis 1979, Woodward 2005). In these worlds significant interfering factors like natural disasters don't occur.

In virtue of the fact that LE supports these interventionist counterfactuals (when it comes to the populations that it models well) it follows directly, on the account of Woodward and Hitchcock (2003a,b), that the variables on the RHS of LE causally explain the actual growth rate of the population $\frac{dn}{dt}$. So, on their account, the cp nature of LE doesn't stand in the way of its playing important aspects of the law role.[4]

3. Minutis Rectis Laws

Not all exceptions to scientific generalizations arise due to the non-fulfilment of (explicit or implicit) cp clauses. This is best illustrated w.r.t. a law that admits of exceptions, but that plausibly is not a cp law, viz. the *Second Law of Thermodynamics* (SLT), which states that the total entropy of an isolated system increases over time, until equilibrium is reached, after which it doesn't decrease.

SLT admits of possible exceptions. Given an initial non-equilibrium state of an isolated system, it is *possible* (though very 'unlikely') that the micro-state should be one that leads to a later state that is further from equilibrium. An example of SLT-violation, which is nevertheless possible (i.e. consistent with the fundamental dynamical laws), is an isolated

[3]Woodward (2005) gives a precise definition of the technical notion of an 'intervention'. For present purposes, it will suffice to think of interventions as ideal experimental manipulations of variables.

[4]The same is true on the accounts given by Lepore and Loewer (1987) and Fodor (1989, 1991), though I focus on Woodward and Hitchcock's account here.

system comprising an ice cube in hot water, in which the ice cube grows larger and colder, while the surrounding water becomes hotter.

Such exceptions to SLT *do not* arise due to failures of a cp condition to hold. SLT is not aptly construed as a cp law. Rather than a cp clause, SLT includes a precise specification of its scope of application: it applies to thermodynamically isolated systems (including the universe as a whole). Unlike LE, there's no possibility of interference from outside the systems that SLT characterizes.

Perhaps the claim that SLT is not a cp law can be disputed. Someone might, for instance, attempt to construe its appeal to an ideal isolated system as somehow amounting to a cp clause (compare Schurz 2002, 369–370). I don't need to insist that it's not a cp law. What I *do* wish to insist is that there is a type of possible exception to it that has nothing to do with the violation of any cp clause. That is, there is a class of exception that is not due to the failure of its idealizations to hold. *Even assuming an ideal isolated system*, exceptions to SLT may arise just as a consequence of certain unlikely microphysical realizations of the system's initial thermodynamic state.[5]

Laws that admit of this type of exception are what I am calling '*minutis rectis (mr)*' laws: that is, laws that hold only when the properties that they concern are realized in the right way. SLT holds only *minutis rectis* because the macro-states that it concerns are multiply realizable by points in the underlying phase space. In a non-equilibrium system, the majority of points in that space (measure ≈ 1) are on entropy-increasing trajectories. However, there are a very few (measure ≈ 0) that are on entropy-decreasing trajectories. SLT only holds if the initial macro-state of an isolated system is realized 'in the right way' – viz. by one of the 'usual' points in phase space that is on a non-entropy-decreasing trajectory.

Though I have illustrated the distinction between the notion of a cp law and that of an mr law w.r.t. a law that's an mr law but plausibly *isn't* a cp law – namely SLT – many special science generalizations hold *both* only cp *and* only mr. Such laws admit of exceptions

[5]It would be inapt to construe SLT as including an implicit cp condition that supposes away such microphysical realizations. That would be to construe SLT's implicit form as something like 'the total entropy of an isolated system is non-decreasing over time, except when the initial microstate is such that it is decreasing'. But this comes close to rendering SLT empty when clearly it isn't (compare Earman and Roberts 1999, 465).

even when their cp clauses are satisfied. Even when there is no disruptive interference from outside the systems that such generalizations characterize (so that their cp conditions are satisfied), they may still be violated just as a consequence of the properties that they concern being realized in the 'wrong' way.

LE is an example of a cp generalization that also holds only mr. I have already argued that it holds only cp. Rather trivially it also holds only mr. LE will break down if members of a population to which it normally applies start *en masse* to exhibit SLT-violating behavior: for example, if neurotransmitters suddenly stop diffusing across their synapses, or oxygen stops diffusing in their blood streams. In such a case, the growth rate of the population will not be predicted by LE. Not for nothing does Lawton (1999, 178) say that SLT is one of the "three deep universal laws that underpin all ecological systems"!

There may also be more interesting reasons why LE holds only mr. For example, the *geographical distribution* of a population can make a difference to its actual growth rate. Indeed, given that population growth can be extremely sensitive to precise initial conditions (see, e.g., May 1974), even very small perturbations of the precise, individual-by-individual initial geographical distribution of members of a population can potentially make a difference to whether the population grows according to LE or sharply declines (even where the population is well below the carrying capacity). The latter situation – in which the population is initially precisely distributed in one of those rare ways that leads to dramatically LE-violating behavior – would be analogous to a thermodynamic system's being at one of those rare points in phase space that leads to SLT-violating behavior. A population's having a certain size, n, is multiply realizable by different precise individual-by-individual geographical distributions. Only if the geographical distribution is 'right' will LE approximately hold.

4. Why It Matters

The distinction between cp and mr generalizations matters because the mr nature of a generalization poses problems for its ability to support counterfactuals and causal-explanatory relations in a way that its cp nature does not.

Consider Woodward and Hitchcock's claim that generalizations like LE are invariant (i.e., support counterfactuals about what would happen) under interventions. The argument that this is so rests upon the idea that the closest worlds in which we intervene upon (say) the population size are not worlds in which the cp condition is violated: in such worlds there is (e.g.) no natural disaster that wipes out the population immediately after the intervention. So, post-intervention, the growth rate is modeled by LE. Given Woodward's notion of an intervention (Woodward 2005, 98) and Lewis's suggested similarity measure over possible worlds (Lewis 1979, 472), this all seems plausible.

Yet the mr nature of LE appears to undercut its ability to support interventionist counterfactuals. Even concerning a population that is usually well-modeled by LE, it seems extremely doubtful that it is true that 'If the size of the population had been intervened upon, the post-intervention micro-state wouldn't have been one that leads to entropy-decreasing behavior'. After all, it seems that the post-intervention micro-state just *might* have been one of those rare entropy-decreasing ones.

It is also doubtful that, even where a population size is well below the carrying capacity, it is true that 'If the size of the population had been intervened upon, the resulting precise geographical distribution would not have been such as to lead to a severe decline in the population'. After all, it's not possible (even metaphysically speaking) to intervene on the size of the population without impacting on the precise individual-by-individual distribution (fewer or more individuals can't be distributed in the same individual-by-individual way) and, in light of the dramatic effects that slight changes in initial conditions can have on ecosystems, the post-intervention distribution just *might* have been one of those rare ones that leads to a dramatic decline in numbers.[6]

It is very doubtful that the truth of either of the counterfactuals considered in the previous two paragraphs follows from the Woodwardian notion of an intervention or the Lewisian notion of similarity among possible worlds. But if such counterfactuals aren't true, then it

[6]If we build into the antecedent of the counterfactual a specification of exactly *how* the intervention would occur (and what the resulting precise geographical distribution would be), then we might get a true counterfactual. But this is not the sort of interventionist counterfactual to which Woodward and Hitchcock appeal in their account of causal explanation.

appears that we can't reason that, if the population size had been intervened upon, then the growth rate would have subsequently followed LE.

Likewise with SLT. Consider the counterfactual 'If I had placed this ice cube into that glass of hot water, then it would have melted quickly'. SLT's mr nature appears to undercut its ability to support this counterfactual. We can't (it seems) say that if the ice cube had been placed in the hot water, then the resulting system *would not* have been in one of those rare micro-states that fails to lead to melting. The post-intervention system *might* have been in such a micro-state, and this undercuts the assertion that the ice cube would have melted. The Lewisian notion of similarity doesn't appear to make a world in which the specified post-intervention macro-state is realized by a non-entropy-increasing micro-state *more dissimilar* to the actual world than one in which it is realized by an entropy-increasing micro-state (compare Hájek (ms)).

If mr laws aren't able to sustain such counterfactuals about what would happen under interventions (i.e. if they're not invariant generalizations), then this threatens to undermine their ability to underwrite causal-explanatory relations and their predictive power. This indicates that philosophical vindications – such as Woodward and Hitchcock's – of the causal-explanatory and predictive power of cp laws are not *ipso facto* vindications of mr laws.

5. Potential Solutions

There is a range of approaches that one might take in attempting to address the problems posed by mr laws.

First, one might consider modifying the Lewisian similarity metric so that worlds in which (e.g.) I intervene on a thermodynamic system and the post-intervention system conforms to SLT come out closer than those in which the post-intervention system does not so conform. We might similarly take conformity to special science laws, like LE, to make for similarity to the actual world.

For example, in response to a worry raised by Elga (2001) about whether Lewis's similarity metric delivers an asymmetry of counterfactual dependence, Dunn (2011) suggests modifying

Lewis's metric so that, other things being equal, worlds obeying SLT and also the various special science laws come out closer to the actual world than those that don't. Such a proposal would seem to ensure that counterfactuals like 'If I had placed the ice cube in the hot water, then it would have melted quickly' come out true, so that SLT supports the counterfactuals needed to underwrite causal/explanatory and predictive relations after all.

One concern about this approach is that it appears to force upon us the truth of counterfactuals like 'If I had put the ice cube in the hot water, then the resulting system wouldn't have been in one of the rare entropy-decreasing microstates'. This counterfactual seems less plausible. But perhaps there is some room for maneuver: perhaps, for instance, one could maintain that the assertion of the latter counterfactual results in a context shift and a corresponding change in the standards of similarity (compare Lewis 1979), with the consequence that this second counterfactual utterance asserts a false proposition (while, in the original, ordinary context, the first asserted a true one). I shan't explore the prospects for such a response here.

A *second* option might be to modify the Woodwardian notion of an intervention so that (e.g.) manipulations of a population size that result in the population being geographically distributed 'in the wrong way', don't count as 'interventions' in the relevant, technical sense. A worry about this strategy is that it is not clear that the 'wrong' sort of interventions could be specified in a systematic and non-ad-hoc way. Simply specifying the relevant 'interventions' in terms of precisely those counterfactual outcomes that one wants to avoid (as in 'the "intervention" on population size must not be such as to result in a precise geographical distribution that leads to a violation of LE') is ad hoc and unsystematic. Similarly, in the thermodynamic case, one might wonder whether there is a useful notion of 'intervention' such that manipulations of a system's macro-state that happen to result in its being in a micro-state on an entropy-decreasing trajectory fail to count as interventions. I shan't explore this strategy further here.

A *third* option would be to argue that 'deterministic' mr laws are mere approximations to probabilistic laws. For example, it is tempting to say that, while SLT is a mr law, it is

an approximation to a probabilistic law that is not a mr law. Statistical mechanics (SM) furnishes us with an exceptionless, *probabilistic* version of SLT.

In SM probabilities are generated by applying a uniform probability distribution (on the Lebesgue measure) to the region of phase space associated with the initial macro-state of an isolated system. Since the measure of points in this phase-space on non-entropy-decreasing trajectories is extremely high, and the measure of points on entropy-decreasing trajectories is extremely low, the result of applying the uniform distribution is an *overwhelmingly high probability* that entropy does not decrease over time (compare Albert 2000, Loewer 2001, Frigg and Hoefer 2014).

Although the probabilistic version of SLT implied by SM *does not* support counterfactuals like 'If I had placed the ice cube in the glass of hot water, then it would have melted quickly', it does support counterfactuals like 'If I had placed the ice cube in the hot water, then *the probability* that it would have melted quickly *would have been very much higher* than if (say) I had returned the ice cube to the freezer'. This appears to be precisely the sort of counterfactual that is relevant to *probabilistic* prediction, causation, and explanation.

I'm sympathetic to this proposal. It is worth noting, however, that this line of response involves construing the probabilities of SM *objectively*. Otherwise, it's hard to see how they could underwrite objective relations of probabilistic causation and explanation. Yet the view that the probabilities of SM are objective has become popular (see, e.g., Albert 2000, 2012, Frigg and Hoefer 2014, Loewer 2001, 2012).

Perhaps similar reasoning can be applied to other high-level laws – like LE – that appear to hold only mr. It might be argued that they too are merely approximations to probabilistic laws that *don't* hold only mr. Rather ambitiously, Albert (2000, 2012) and Loewer (2001, 2008, 2012) argue that SM itself actually entails probabilistic approximations of the laws of the special sciences. While this 'Statistical Mechanical Imperialism' has been criticized (Callender 2011, Frisch 2014, Weslake 2014), one might nevertheless think that there *are* probabilistic approximations to special science mr laws (perhaps derivable in some other

way). If so, then these probabilistic laws may be able to support the counterfactuals relevant to probabilistic causal explanation and prediction.

For instance, in ecology the geographical distribution of populations is often modeled via a probability distribution (e.g. the Poisson distribution) over a habitat (see, e.g., Vandermeer and Goldberg 2013, 126-142). Perhaps, in general, we can get strict(er) probabilistic versions of special science generalizations via the imposition of probability distributions over the underlying state-spaces in which the properties that they concern are realized.

A lot of work needs to be done to show that this will work out. It's reasonable to wonder whether we can *always* replace deterministic special science generalizations that hold only mr with strict(er) probabilistic laws by imposing probability distributions over underlying state spaces. It's also reasonable to wonder whether we can *always* interpret the resulting probabilities objectively. That these are serious questions for both science and metaphysics shows the depth of the challenges posed by the mr nature of many special science generalizations.

6. Conclusion

The notion of a non-exceptionless law shouldn't be equated with that of a cp law. There is another important category of non-exceptionless law that ought to be distinguished, viz. mr laws. The mr nature of special science generalizations poses distinctive challenges for those aiming to show that special science generalizations can support counterfactuals, causal explanations, and predictions. Distinguishing the two categories of non-exceptionless law brings these challenges to light, but also allows us to identify possible avenues for addressing them.

Acknowledgments

Thanks to audiences at the Universities of Cologne, Wuppertal, Munich, and Cambridge, and at the 2013 BSPS Conference. I acknowledge the support of the Deutsche Forschungsgemeinschaft (Grant: SP279/15-1), the McDonnell Causal Learning Collaborative, and the Alexander von Humboldt Foundation.

References

Albert, D. (2000). *Time and chance*. Cambridge, MA: Harvard University Press.

Albert, D. (2012). Physics and chance. In Y. Ben-Menahem and M. Hemmo (Eds.), *Probability in Physics*, pp. 17–40. Berlin: Springer.

Callender, C. (2011). The past histories of molecules. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 83–113. New York: OUP.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: OUP.

Colyvan, M. and L. Ginzburg (2003). Laws of nature and laws of ecology. *Oikos 101*, 649–653.

Davidson, D. (1970). Mental events. In L. Foster and J. Swanson (Eds.), *Experience and Theory*, pp. 79–101. Amherst: University of Massachusetts Press.

Dunn, J. (2011). Fried eggs, thermodynamics, and the special sciences. *The British Journal for the Philosophy of Science 62*, 71–98.

Earman, J. and J. Roberts (1999). *Ceteris paribus*, there is no problem of provisos. *Synthese 118*, 439–478.

Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science 68*, 313–324.

Fodor, J. (1989). Making mind matter more. *Philosophical Topics 17*, 59–79.

Fodor, J. (1991). You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanations. *Mind 100*, 19–34.

Frigg, R. and C. Hoefer (2014). The Best Humean System for statistical mechanics. Forthcoming in *Erkenntnis*.

Frisch, M. (2014). Why physics can't explain everything. In A. Wilson (Ed.), *Asymmetries of Chance and Time*. Oxford: OUP.

Hájek, A. (ms.). Most counterfactuals are false.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy 98*, 273–299.

Lawton, J. (1999). Are there general laws in ecology? *Oikos 84*, 177–192.

Lepore, E. and B. Loewer (1987). Mind matters. *Journal of Philosophy 84*, 630–642.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*, 455–476.

Loewer, B. (2001). Determinism and chance. *Studies in History and Philosophy of Science Part B 32*, 609–620.

Loewer, B. (2008). Why there is anything except physics. In J. Hohwy and J. Kallestrup (Eds.), *Being Reduced*, pp. 149–163. Oxford: OUP.

Loewer, B. (2012). The emergence of time's arrows and special science laws from physics. *Interface focus 2*, 13–19.

May, R. (1974). Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science 186*, 645–647.

Reutlinger, A., G. Schurz, and A. Hütemann (2011). Ceteris paribus laws. *The Stanford Encyclopedia of Philosophy*. URL: $\langle http://plato.stanford.edu/entries/ceteris-paribus/\rangle$.

Schrenk, M. (2007). Can capacities rescue us from ceteris paribus laws? In M. Kistler and B. Gnassounou (Eds.), *Dispositions in Philosophy and Science*, pp. 221–247. Aldershot: Ashgate.

Schurz, G. (2002). Ceteris paribus laws: Classification and deconstruction. *Erkenntnis 57*, 351–372.

Tsoularis, A. and J. Wallace (2002). Analysis of logistic growth models. *Mathematical Biosciences 179*, 21–55.

Turchin, P. (2001). Does population ecology have general laws? *Oikos 94*, 17–26.

Vandermeer, J. and D. Goldberg (2013). *Population Ecology* (2^{nd} ed.). Princeton, NJ: Princeton University Press.

Weslake, B. (2014). Statistical mechanical imperialism. In A. Wilson (Ed.), *Asymmetries of Chance and Time*. Oxford: OUP.

Woodward, J. (2002). There is no such thing as a *ceteris paribus* law. *Erkenntnis 57*, 303–328.

Woodward, J. (2005). *Making Things Happen*. New York: OUP.

Woodward, J. and C. Hitchcock (2003a). Explanatory generalizations, Part I: A counterfactual account. *Noûs 37*, 1–24.

Woodward, J. and C. Hitchcock (2003b). Explanatory generalizations, Part II: Plumbing explanatory depth. *Noûs 37*, 181–199.

A pluralist view about information

Sebastian Fortin[1] – Olimpia Lombardi[1] – Leonardo Vanni[2]

[1] CONICET – Universidad de Buenos Aires
[2] Universidad de Buenos Aires

Abstract

Focusing on Shannon information, this article shows that, even on the basis of the same formalism, there may be different interpretations of the concept of information, and that disagreements may be deep enough to lead to very different conclusions about the informational characterization of certain physical situations. On this basis, a pluralist view is argued for, according to which the concept of information is primarily a formal concept that can adopt different interpretations that are not mutually exclusive, but each useful in a different specific context.

1. Introduction

In the Book 11 of his *Confessions*, St. Augustine asks himself: "What, then, is time? If no one asks me, I know what it is. But if I wish to explain it to one that asketh, I know not." Something similar happens today with information. Both in everyday life and in science, the word 'information' is so pervasive that we all believe we know what we mean by it. However, as soon as we are asked for its precise meaning, the opinions substantially diverge.

As many recognize, information is a polysemantic concept that can be associated with different phenomena (Floridi 2010). In this conceptual tangle, the first distinction to be introduced is between a semantic and a non-semantic view of information. According to the first view, information is something that carries semantic content (Bar-Hillel and Carnap 1953; Bar-Hillel 1964), and which is therefore strongly related with semantic notions such as reference, meaning and representation. In general, semantic information is carried by propositions that intend to represent states of affairs; so, it has "aboutness", that is, it is directed to other things. And although it is still controversial whether false factual content may qualify as information, semantic information is strongly linked with the notion of truth.

Non-semantic information, also called 'mathematical' or 'statistical', is concerned with the statistical properties of a given system and/or the correlations between the states of two systems, independently of the meanings of those states. The classical *locus* of mathematical information is the paper where Shannon (1948) introduces a precise formalism designed to solve certain specific technological problems. Shannon's theory is purely quantitative, it ignores any issue related to informational content: "[the] *semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages*" (Shannon 1948, 379).

Although Shannon's theory is the traditional formalism to quantify information, it is not the only one. For instance, Fisher information measures the dependence of a random variable X on an unknown parameter θ upon which the probability of X depends (Fisher 1925), and algorithmic information measures the length of the shortest program that produces a string on a universal Turing machine (Chaitin 1987). In quantum information theory, von Neumann entropy gives a measure of the quantum resources necessary to faithfully encode the state of the source-system (Schumacher 1995).

It might be supposed that, when confined to a formal framework, the meaning of 'information' is clear: given the mathematical theory, information is what this theory describes. However, this is not the case. Even on the basis of the same formalism, there may be different interpretations of the concept of information, and disagreements may be deep enough to lead to different conclusions in certain physical situations. Although disagreements may arise regarding any formalism, we will focus on Shannon's theory since it is the most widespread formalism, even applicable in the quantum context (Rovelli 1996; Timpson 2003). Finally, we will argue for a pluralist view according to which, once mathematically characterized, the concept of information is a formal concept that can adopt different interpretations not mutually exclusive, each useful in a different context.

2. Shannon's Theory

According to Shannon's theory (Shannon 1948), transmission of information requires a source *S*, a receiver *R* and a channel *CH*. If *S* has a range of possible states $s_1,...,s_n$ –letters–, whose respective probabilities of occurrence are $p(s_1),...,p(s_n)$, the amount of information generated at the source by the occurrence of s_i is defined as $I(s_i) = \log(1/p(s_i))$. When 'log' is the logarithm to the base 2, the resulting unit of measurement is called '*bit*' (if the natural logarithm is used, the unit is the *nat*, and in

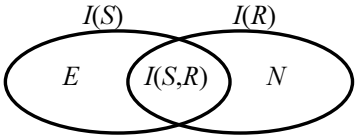
the case of the logarithm to the base 10, the unit is the *Hartley*). Since S produces long sequences of states –messages–, the average amount of information generated at the source is defined as:

$$I(S) = \sum_{i=1}^{n} p(s_i)\ \log(1/p(s_i))$$

Analogously, if the possible states of R are $r_1,...,r_m$, with respective probabilities $p(r_1),...,p(r_m)$, the amount of information received at the receiver by the occurrence of r_j is $I(r_j) = \log(1/p(r_j))$, and the average amount of information received at the receiver is:

$$I(R) = \sum_{j=1}^{m} p(r_j)\ \log(1/p(r_j))$$

The relationship between $I(S)$ and $I(R)$ can be represented as:



$I(S;R)$: mutual information
E : equivocation
N : noise

where $I(S;R) = I(S) - E = I(R) - N$ is the information generated at S and received at R, E is the information generated at S but not received at R, and N is the information received at R but not generated at S (always average amounts). E and N are measures of the dependence between S and R and, therefore, are functions not only of S and R, but also of the channel CH, defined by the matrix $\left[p(r_j/s_i)\right]$, where $p(r_j/s_i)$ is the conditional probability of the occurrence of r_j given the occurrence of s_i, and the elements in any row must sum to 1. Thus, N and E are computed as:

$$N = \sum_{i=1}^{n} p(s_i) \sum_{j=1}^{m} p(r_j/s_i) \log(1/p(r_j/s_i)) \qquad E = \sum_{j=1}^{m} p(r_j) \sum_{i=1}^{n} p(s_i/r_j) \log(1/p(s_i/r_j))$$

where $p(s_i/r_j) = p(r_j/s_i)p(s_i)/p(r_j)$.

One of the most relevant results in Shannon's theory is the noiseless coding theorem, according to which the value of $I(S)$ is equal to the average number of bits necessary to code a letter of the source using an ideal code: $I(S)$ measures the optimal compression of the source messages. In fact, the messages of N letters produced by S fall into two classes: one of approximately $2^{NI(S)}$ typical messages, and the other of atypical messages. When $N \to \infty$, the probability of an atypical message becomes negligible; so, the source can be conceived as producing only $2^{NI(S)}$ possible messages. This suggests a

natural strategy for coding: each typical message is coded by a binary sequence of length $NI(S)$, in general shorter than the length N of the original message.

Given this formalism, it seems that there is nothing controversial in the concept of Shannon information: it would be what Shannon's theory describes. However, matters are not so simple.

During the last years, it has been usual to hear in the philosophy of physics (not in the physics) community that the problem of the interpretation of information is dissolved because the word 'information' is an abstract noun. Timpson (2004, 2008) insists that what is produced at the source and that we desire to transmit is not a *token*-sequence but a *type*-sequence; but types are abstract and, so, they are not part of the spatio-temporal content of the world. Therefore, according to this view information is not a substance, not even a physical entity, because it is not an entity at all: there is nothing the word 'information' refers to.

Despite the diffusion of this position, one may suspect that information is even more abstract than a type. In fact, types are not items to be measured in bits. Moreover, the information of very different types may be the same, since the only relevant aspect in Shannon information is that the actual sequence is one selected from a set of possible sequences. And it is not even the case that we always want to transmit the same type-sequence: the states of the receiver may be completely different, even in a type sense, than the states of the source: the success of information transmission depends on the decision about the expected correlations, embodied in the fidelity function, between the source states and the receiver states. In brief, Timpson unwittingly reintroduces semantic issues –analogous to those related with the difference between proposition, sentence and utterance– in the discussion about Shannon information, a field where semantics plays no role at all.

Of course, these briefs comments are not a full analysis of Timpson's very articulated position, which deserves a specific article. Nevertheless, they open the way to focus on the different views about Shannon information that are still present in philosophical and physical discussions.

3. Epistemic and Physical Interpretations of Information

A concept usually connected with the notion of information is that of knowledge: information provides knowledge, modifies the state of knowledge of those who receive it. Some believe that the link between information and knowledge is a feature of the everyday notion of information, which must be carefully distinguished from the Shannon's technical concept (Timpson 2004). However, the idea of knowledge is present also in the philosophical and the physical discourse about information.

In fact, it is common to find authors who even define information in terms of knowledge. For instance, on the basis of Shannon's theory as the underlying formalism for his proposal, Dretske says: "*information is a commodity that, given the right recipient, is capable of yielding knowledge.*" (1981, 47). According to MacKay, information is linked to an increase in knowledge on the receiver's side: "*Suppose we begin by asking ourselves what we mean by information. Roughly speaking, we say that we have gained information when we know something now that we didn't know before; when 'what we know' has changed*." (1969, 10).

This presence of the notion of knowledge is not confined to authors who try to supply a semantic content to statistical information. Some philosophers of physics are also persuaded that the core meaning of the concept of information, even in its technical sense, is linked to the concept of knowledge (Myrvold, personal communication). And physicists frequently speak about what we know or may know when dealing with information. For instance, Rovelli (1997) insists that quantum mechanics is a theory about information because it talks about the relations between what different observers "know" about a quantum system. Zeilinger even equates information and knowledge when he says that "*We have knowledge, i.e., information, of an object only through observation*" (1999, 633) or, with Bruckner, "*For convenience we will use here not a measure of information or knowledge, but rather its opposite, a measure of uncertainty or entropy*" (2009, 681-82). Even in a traditional textbook about Shannon's theory one can read that information "*is measured as a difference between the state of knowledge of the recipient before and after the communication of information.*" (Bell 1957, 7), and that it must be relativized with respect to the background knowledge available before the transmission: "*the datum point of information is then the whole body of knowledge possessed at the receiving end before the communication*." (Bell 1957, 7).

It is worth stressing that, from the epistemic perspective, the possibility of acquiring knowledge about a source by consulting the state of a receiver is rooted in the nomic character of the regularities underlying the whole situation. In fact, the conditional probabilities that define the channel do not represent merely *de facto* correlations; they are determined by a network of lawful connections between the states of the source and the states of the receiver.

A different view about information is the one that detaches the concept from the notion of knowledge and considers information as a physical magnitude. This is the position of many physicists and most engineers, for whom the essential feature of information is its capacity to be generated at one point of the physical space and transmitted to another point; it can also be accumulated, stored and

converted from one form to another, like other physical magnitudes such as energy. In this case, the capability of providing knowledge is not a central issue, since the transmission of information can be used only for control purposes, such as controlling a device at the receiver end by modifying the state of the source. According to this view, it is precisely because of the physical nature of information that the dynamics of its flow is constrained by physical laws and facts: "*Information handling is limited by the laws of physics and the number of parts available in the universe*" (Landauer 1991, 29; see also Bennett and Landauer 1985).

In general, the physical interpretation of information comes strongly linked with the idea expressed by the well-known *dictum* 'no information without representation': the transmission of information between two points of physical space necessarily requires an information-bearing signal, that is, a physical process propagating from one point to the other. Landauer is an explicit defender of this position when he claims that "*Information is not a disembodied abstract entity; it is always tied to a physical representation. It is represented by engraving on a stone tablet, a spin, a charge, a hole in a punched card, a mark on a paper, or some other equivalent*." (1996, 188). This view is also adopted by some philosophers of science; for instance, Kosso states that "*information is transferred between states through interaction.*" (1989, 37). The need of a carrier signal is natural in the light of the generic idea that physical influences can only be transferred through interactions.

In the context of this physical interpretation, information tends to be compared with energy, which was born in the specific field of mechanics as a pragmatic notion related with the resources we can draw from a mechanical system, but ended up being conceived as a highly wide reaching concept: at present the word 'energy' refers to an item that pervades the whole world of physics. On this basis, information is conceived by many physicists as a physical entity with the same ontological status as energy; it has also been claimed that its essential property is the power to manifest itself as structure when added to matter (Stonier 1990, 1996).

4. Epistemic versus Physical Interpretations of Information

If the difference between the epistemic and the physical interpretations of information is clear from a conceptual viewpoint, it turns out to be even more clear when the concept of information is applied to particular situations.

Let us consider a source S that transmits information to two physically isolated receivers R_A and R_B via a certain physical link. In this case, the correlations between the states of the two receivers are

not accidental, but functions of the physical dependence of R_A and R_B on S. Nevertheless, there is no physical interaction between the receivers. The informational description of this situation is completely different from the viewpoints given by the two interpretations of the concept of information. According to the physical interpretation, it is clear that no information is being transferred between R_A and R_B since there is no physical signal traveling between them. However, from an epistemic interpretation, nothing prevents us from admitting the existence of an informational link between the two receivers. In fact, we can define a communication channel between R_A and R_B because it is possible to learn something about R_B by looking at R_A and vice versa: "*from a theoretical point of view* [. . .] *the communication channel may be thought of as simply the set of depending relations between* [a system] *S and* [a system] *R. If the statistical relations defining equivocation and noise between S and R are appropriate, then there is a channel between these two points, and information passes between them, even if there is no direct physical link joining S with R.*" (Dretske 1981, 38). The receiver R_B may even be farther from the source S than R_A, so that the events at R_B may occur later than those at R_A. Nevertheless, this is irrelevant from the epistemic view of information: although the events at R_B occur later, R_A carries information about what will happen at R_B.

Somebody might consider that the difference in the informational characterization of the situation described above is a mere curiosity with no philosophical interest. However, this kind of disagreements has also relevant consequences in the characterization of central notions in the philosophy of science. For instance, there is an important philosophical tradition that explains scientific observation in terms of information. In order to elucidate the notion of observation without resorting to perceptual matters, Shapere proposes that x is directly observed if information is received by an appropriate receptor and that information is transmitted from the entity x to the receptor without interference (Shapere 1982). Brown agrees with Shapere in stressing that observing an item I consists in gaining information about I by examining another item I^* (Brown 1987). Kosso (1989) also adheres to this tradition with his "interaction-information" account of scientific observation.

In general (with the exception of Kosso, who relies on Shannon's theory), in the discussions about scientific observation the concept of information is not sufficiently specified in formal terms, so the interpretation of the concept is even less considered. However, the way in which information is conceived leads to very different consequences regarding the view about observation. This turns out to be particularly clear in the so-called 'negative experiments', which were originally devised as a theoretical tool for analyzing the quantum measurement problem (see Jammer 1974). Nevertheless,

they can be regarded independently from quantum mechanics: in a negative experiment it is assumed that an event has been observed by noting the absence of some other event. This is the case of neutral weak currents, which are observed by noticing the absence of charged muons. But the conceptual core of negative experiments can be understood by means of a very simple example. Let us consider a tube in whose middle point a classical particle is emitted towards one of the ends of the tube; a detection device is placed at one of the ends, say *A*, in order to know in which direction the particle was emitted. Since there is a perfect anticorrelation between both ends of the tube, by looking at the right end *A*, we can know the state at the left end *B*. Nevertheless, the instantaneous propagation of a signal between *A* and *B* is physically impossible. If after an appropriate time –depending on the velocity of the particle and the length of the tube– the device at *A* indicates no detection, we can conclude that the particle was emitted toward the left end *B*. But, have we observed the direction of the emitted particle? From an informational account of scientific observation, the answer depends on the interpretation of the concept of information adopted. On the basis of an epistemic interpretation, a communication channel between the two ends of the tube can be defined, which allows us to *observe* the presence of the particle at *B*, even though there is no signal between *B* to *A*. The physical view leads us to a concept of observation narrower than the previous one: by looking at the detector we observe the state at *A*, but we do not observe the state at *B*; such a state is *inferred.*

As it has been repeatedly noticed, Shannon information is not tied to classical physics: any type of physical system can be used to design the informational situation (Timpson 2003, 2004; Dwell 2003). Therefore, Shannon's theory can in principle be applied to the quantum domain, in particular, to EPR-type experiments, characterized by theoretically well-founded correlations between two spatially separated particles. During many years it was repeated that information cannot be sent between both particles because the propagation of a superluminal signal from one particle to the other is impossible: there is no information-bearing signal that can be modified at one point of space in order to carry information to the other spatially separated point. But the fact that the physical interpretation of information underlies that claim was usually not noticed. On the contrary, the epistemic interpretation, which only requires correlations, would face no problem in defining an informational channel between the two EPR-particles.

Disagreements increase when quantum information comes into play. Teleportation is one of the paradigmatic phenomena in this field. Broadly speaking, an unknown quantum state is transferred from Alice to Bob with the assistance of a shared pair prepared in an entangled state and of two classical bits

sent from Alice to Bob (the description of the protocol can be found in any text on the matter). Although the situation is usually not strictly described in informational terms (not Shannon's nor quantum informational terms), the idea is that the very large (strictly infinite) amount of information required to specify the teletransported state is transferred from Alice to Bob by sending only two bits. When addressing this problem, many physicists try to find a physical link between Alice and Bob that could play the role of carrier of information. For instance, Penrose (1998) and Jozsa (1998, 2004) claim that information may travel backwards in time: "*How is it that the continuous 'information' of the spin direction of the state that she wishes to transmit* […] *can be transmitted to Bob when she actually sends him only two bits of discrete information? The only other link between Alice and Bob is the quantum link that the entangled pair provides. In spacetime terms this link extends back into the past from Alice to the event at which the entangled pair was produced, and then it extends forward into the future to the event where Bob performs his*." (Penrose 1998, 1928). According to Deutsch and Hayden (2000), the information travels hidden in the classical bits. These physicists do not explicitly acknowledge that the problem derives from the physical interpretation of information to which they strongly adhere, and that an epistemic view would not commit them to find a physical channel between Alice and Bob.

Of course, an elucidation of the concept of information does not dissolve all the conundrums involved in teleportation (see Timpson 2006), or in the phenomenon of entanglement that underlies it. Nevertheless, such elucidation would help us to find a way out of the problems derived from the informational characterization of teleportation. One may wonder how essential the need of a spatio-temporal link is in the physical interpretation of information. Or one may reconstruct the situation in Shannon terms to conclude that the information effectively transmitted (the mutual information) is really not very large, to the extent that the receiver cannot retrieve the whole information generated at the source. Or one may even decide to leave aside the physical interpretation in favor of an epistemic view that recovers the relation between information and knowledge.

5. A Pluralist Approach to Information

Up to this point, the epistemic and the physical interpretations of Shannon information were presented as rival; nevertheless, this is not necessarily the case.

Although the physical interpretation has been the most usual in the traditional textbooks used in engineer's training, this has changed in recent times: in general, present-day textbooks explain information theory in a formal way, with no mention of sources, receivers or signals, and the basic

concepts are introduced in terms of random variables and probability distributions over their possible values. Only when the formalism has been presented, is the theory applied to the traditional case of communication. For instance, in their extensively used book Cover and Thomas emphasize that: "*Information theory answers two fundamental questions in communication theory* […]. *For this reason some consider information theory to be a subset of communication theory. We will argue that it is much more. Indeed, it has fundamental contributions to make in statistical physics* […]*, computer sciences* […]*, statistical inference* […] *and to probability and statistics.*" (1991, 1)

The idea that the concept of information is completely formal is not new. Already Khinchin (1957) and Reza (1961) conceived information theory as a new chapter of the theory of probability. From this perspective, Shannon information not only is not a physical magnitude, but it also loses its nomic ingredient: the mutual information between two random variables can be defined even if there is no lawful relationship between them and their conditional probabilities express only *de facto* correlations.

If the concept of information is purely formal and belongs to a mathematical theory, the word 'information' does not belong to the language of empirical sciences –or to ordinary language–: it has no extralinguistic reference in itself. Its "meaning" has only a syntactic dimension. According to this view, the generality of the concept of Shannon information derives from its exclusively formal nature; this generality is what makes it a powerful formal tool for empirical science, applicable to a variety of fields.

From this formal perspective, the relationship between the word 'information' and the different views of information is the logical relationship between a mathematical object and its interpretations, each one of which endows the term with a specific referential content. The epistemic view, then, is only one of the different possible interpretations, which may be applied in psychology and in cognitive sciences by using the concept of information to conceptualize the human abilities of acquiring knowledge (see e.g. Hoel, Albantakis and Tononi 2013). The epistemic interpretation might also serve as a basis for the philosophically motivated attempts to add a semantic dimension to a formal theory of information (MacKay 1969; Nauta 1972; Dretske 1981)

In turn, the physical view, which makes information a physical magnitude carried by signals, is clearly the appropriate interpretation in communication theory, in which the main problem consists in optimizing the transmission of information by means of physical carriers whose energy and bandwidth is constrained by technological and economic limitations. But this is not the only possible physical

interpretation: if S is not interpreted as a source with states but a macrostate compatible with many equiprobable microstates, $I(S)$ represents the Boltzmann thermodynamic entropy of S. Furthermore, in computer sciences a communicational information may be defined, such that, if S is interpreted as a binary string of finite length, $I(S)$ can be related with the algorithmic complexity of S. The understanding of the relationship between the formal concept of information and its interpretations serves for assessing the usually obscure extrapolations from communication theory to thermodynamics or computing.

Summing up, this pluralist view about information rejects the question about "the" meaning of information: "*The word 'information' has been given different meanings by various writers in the general field of information theory*. [...] *It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field*." (Shannon 1993, 180).

6. References

Bar-Hillel, Yehoshua 1964. *Language and Information: Selected Essays on Their Theory and Application*. Reading, Mass: Addison-Wesley.

Bar-Hillel, Yehoshua, and Rudolf Carnap 1953. "Semantic Information." *The British Journal for the Philosophy of Science* 4:147-57.

Bell, David 1957. *Information Theory and its Engineering Applications*. London: Pitman & Sons.

Bennett, Charles, and Rolf Landauer 1985. "The Fundamental Physical Limits of Computation." *Scientific American* 253:48-56.

Brown, Harold 1987. *Observation and Objectivity*. Oxford: Oxford University Press.

Brukner, Časlav, and Anton Zeilinger 2009. "Information Invariance and Quantum Probabilities." *Foundations of Physics* 39:677-89.

Chaitin, Gregory 1987. *Algorithmic Information Theory*. New York: Cambridge University Press.

Cover, Thomas, and Joy Thomas 1991. *Elements of Information Theory*. New York: JohnWiley & Sons.

Deutsch, David, and Patrick Hayden 2000. "Information Flow in Entangled Quantum Systems." *Proceedings of the Royal Society of London A* 456:1759-74.

Dretske, Fred 1981. *Knowledge and the Flow of Information*. Oxford: Basil Blackwell.

Duwell, Armond 2003. "Quantum Information Does Not Exist." *Studies in History and Philosophy of Modern Physics* 34:479-99.

Fisher, Ronald 1925. “Theory of Statistical Estimation.” *Proceedings of the Cambridge Philosophical Society* 22:700-25.

Floridi, Luciano 2010. *Information – A Very Short Introduction*. Oxford: Oxford University Press.

Hoel, Erik, Larissa Albantakis, and Giulio Tononi 2013. “Quantifying Causal Emergence Shows that Macro Can Beat Micro.” *Proceedings of the National Academy of Sciences* 110:19790-95.

Jammer, Max 1974. *The Philosophy of Quantum Mechanics*. New York: John Wiley & Sons.

Jozsa, Richard 1998. “Entanglement and Quantum Computation.” In *The Geometric Universe*, ed. S. Huggett, L. Mason, K. P. Tod, S. T. Tsou, and N. M. J. Woodhouse, 369-79. Oxford: Oxford University Press.

—— 2004. “Illustrating the Concept of Quantum Information.” *IBM Journal of Research and Development* 4:79-85.

Khinchin, Aleksandr 1957. *Mathematical Foundations of Information Theory*. New York: Dover.

Kosso, Peter 1989. *Observability and Observation in Physical Science*. Dordrecht: Kluwer.

Landauer, Rolf 1991. “Information is Physical.” *Physics Today* 44:23-29.

—— 1996. “The Physical Nature of Information.” *Physics Letters A* 217:188-93.

MacKay, Donald 1969. *Information, Mechanism and Meaning*. Cambridge: MIT Press.

Nauta, Doede 1972. *The Meaning of Information*. The Hague: Mouton.

Penrose, Roger 1998. “Quantum Computation, Entanglement and State Reduction.” *Philosophical Transactions of the Royal Society of London A* 356:1927-39.

Reza, Fazlollah 1961. *Introduction to Information Theory*. New York: McGraw-Hill.

Rovelli, Carlo 1996. “Relational Quantum Mechanics.” *International Journal of Theoretical Physics* 35:1637-78.

Schumacher, Benjamin 1995. “Quantum Coding.” *Physical Review A* 51:2738-47.

Shannon, Claude 1948. “The Mathematical Theory of Communication.” *Bell System Technical Journal* 27:379-423.

—— 1993. *Collected Papers*, ed Neil Sloane, and Aaron Wyner. New York: IEEE Press.

Shapere, Dudley 1982. “The Concept of Observation in Science and Philosophy.” *Philosophy of Science* 49:485-525.

Stonier, Tom 1990. *Information and the Internal Structure of the Universe: An Exploration into Information Physics*. New York-London: Springer.

—— 1996. “Information as a Basic Property of the Universe.” *Biosystems* 38:135-40.

Timpson, Christopher 2003. "On a Supposed Conceptual Inadequacy of the Shannon Information in Quantum Mechanics." *Studies in History and Philosophy of Modern Physics* 34:441-68.

—— 2004. *Quantum Information Theory and the Foundations of Quantum Mechanics*. PhD diss., University of Oxford (quant-ph/0412063).

—— 2006. "The Grammar of Teleportation." *The British Journal for the Philosophy of Science* 57:587-621.

—— 2008. "Philosophical Aspects of Quantum Information Theory." In *The Ashgate Companion to the New Philosophy of Physics*, ed. Dean Rickles, 197-261. Aldershot: Ashgate Publishing.

Zeilinger, Anton 1999. "A Foundational Principle for Quantum Mechanics." *Foundations of Physics* 29:631-43.

SHOULD A HISTORICALLY MOTIVATED ANTI-REALIST BE A STANFORDITE?

ABSTRACT: Suppose one believes that the historical record of discarded scientific theories provides good evidence against scientific realism. Should one adopt Kyle Stanford's specific critique of realism? I present reasons for answering this question in the negative. In particular, Stanford's challenge, based on the problem of unconceived alternatives, cannot use many of the *prima facie* strongest pieces of historical evidence against realism: (i) superseded theories whose successors were explicitly conceived, and (ii) superseded theories that were not the result of elimination-of-alternatives inferences.

1. Introduction.

Scientific realism's opponents have long appealed to the history of science as evidence for their position. The most important recent development in this tradition is probably Kyle Stanford's Problem of Unconceived Alternatives (PUA), which supposedly explains his New Induction (NI) over the history of science. According to the PUA, "our cognitive constitutions or faculties are not well-suited to exhausting the kinds of spaces of serious alternative theoretical possibilities from which our fundamental theories of nature are drawn" (2006, 45). In other words, in 'fundamental' scientific theorizing, scientists lack the cognitive ability to devise all plausible hypotheses that would explain the evidence available to them. The NI states that historical scientists often failed to exhaust the space of scientifically respectable hypotheses that would explain the evidence available to them at that historical time; therefore, present scientists are probably failing similarly. For example, General Relativity can explain all the data that was available to Newton, but Einstein's theory was not conceived until the early 20th Century. Stanford claims this creates a problem for realism, because many scientific theories are inferred via an elimination-of-alternatives inference (also known as 'disjunctive syllogism') (2006, 28). In an eliminative inference, a supposedly exhaustive list of hypotheses ($H_1 \vee H_2 \vee \ldots \vee H_n$) is proposed, and all are eliminated ($\neg H_2, \ldots \neg H_n$) except one ($H_1$); we

conclude the single remaining hypothesis is correct. But the NI gives us reason to believe that, for 'fundamental' domains of scientific theorizing, the list of hypotheses probably does *not* contain a true hypothesis, so the disjunction is probably untrue. Therefore, in those fundamental domains such an argument would be unsound, and thus we lack sufficient evidence to believe scientific theory H_1 is true. Call this the *PUA-based argument against realism*.

Many critics of Stanford's position reject anti-realism. This paper takes a different perspective: suppose one *is* moved by the 'dustbin of history,' and wishes to be a historically motivated anti-realist. Should one accept Stanford's particular version of this view? I present evidence for a negative answer: the PUA-based argument against realism omits much of the best historical evidence against scientific realism, and as a result, delivers an unnecessarily restricted version of anti-realism. In particular, many discarded theories that are *prima facie* strong evidence against realism involve either *conceived* alternatives or non-eliminative ('projective') inferences. Section 2 describes historical examples of such conceived alternatives, and section 3 lists theories that were inferred projectively, but were later discarded. Along the way, I argue that the most natural attempts to accommodate these cases within the PUA either fail on their own terms or contravene Stanford's other commitments.

Now, Stanford might reply that he can accept my claims in §§2-3 without contradiction. I accept that the NI and PUA, considered in isolation, are consistent with the points in §§2-3. However, Section 4 argues that some of Stanford's other central claims *are* in tension with those points. In particular, if a Stanfordite accepts the main contentions of §§2-3, then she must reject Stanford's claims about the importance of the

PUA, and relinquish the existence of an epistemic distinction between projective inferences and eliminative ones—a distinction Stanford needs for his instrumentalism to be piecemeal or selective instead of global.

2. *Conceived* Alternatives

I grant that the NI provides evidence against realism. However, if a historically motivated anti-realist restricts her evidence to cases where the alternatives to the prevailing theory were unconceived, then she omits some of the *prima facie* best historical evidence for anti-realism. Much of the most compelling evidence for historically based anti-realism involves cases where the successor theory was conceived explicitly—and explicitly rejected at that earlier time as inferior to the now-discarded theory.

2.1. Examples

One set of interrelated examples appears in Book I of Ptolemy's *Almagest*. In I.5, Ptolemy argues that the Earth must be at the center of the universe, by assuming for reductio that it is not, and deducing claims that contradict accepted observations (specifically, an observer anywhere on the Earth always sees half of the zodiac). In I.7, he considers the hypothesis that the Earth is moving from one place to another, and argues that it is impossible, given the arguments in I.5 (for if it were moving from one place to another, it could not be at the center all the time). The absence of stellar parallax is further evidence that the earth does not move from one place to another. Later in I.7, Ptolemy considers the hypothesis that the Earth rotates on its axis. He admits that this

hypothesis is consistent with the celestial phenomena, but argues that no version of this hypothesis is consistent with terrestrial phenomena.

These three hypotheses—that the Earth has a translational motion, rotates daily, and is not at the center of the universe—were explicitly conceived by Ptolemy. He rejected each of them because he thought the balance of the evidence told against them. This shows that even when scientists evaluate a hypothesis that later scientists will come to believe superior, the earlier scientists can reject it. The NI cannot appeal to such cases as evidence against realism, since they involve *conceived* alternatives that are later accepted.

These hypotheses are not isolated instances. The hypothesis that the heat of a body consists in the motion of that body's parts was in the same situation from the early-to-mid 1700's until about 1840.[1] The view that heat was the motion of component corpuscles was defended by several luminaries of the Scientific Revolution. However, as the 1700's progressed, this hypothesis came to be regarded as inferior to the view that heat was material by many of the leading researchers at the time.[2] These material caloric theorists had certainly considered the view that heat was motion of the constituent particles, since they had read the mechanical philosophers, but they rejected it.

[1] The cut-off date of 1840 comes from (Brush 1976, 27), but the decline was gradual: (Metcalfe 1859) criticizes the hypothesis that heat is the motion of constituent parts.

[2] Defenders of caloric sometimes admitted that the evidence did not demonstrate that heat was material. Joseph Black is typical: "Such an idea [*viz*., material caloric] of the nature of heat is, therefore, the most probable of any I know; … It is, however, altogether a supposition" (Black 1803, 33). However, Black also says that the mechanical theory is "totally inconsistent with the phenomena" (80).

Enlightenment caloric theorists brought several arguments against the mechanical view of their predecessors (Brush 1976, 28-30); perhaps most compelling was the fact that heat can diffuse across a vacuum, but the motion of particles cannot be transmitted across a space that contains no particles.

There are further examples, such as the mutability of species. Linnaeus, for example, concisely argues that "no new species are produced" (1964 [1735], 18), arguing against the successor hypothesis that new species are created. Several other candidates for further examples are considered in (Hook 2002), a collection on Gunther Stent's notion of a 'premature' hypothesis; (Barber 1961) contains a classic list of hypotheses that were considered, rejected, and later accepted. In sum: restricting focus to the problem of *un*conceived alternatives shrinks the body of evidence available in support of historically-motivated anti-realism—for each of these cases involves *conceived* alternatives to the then-dominant theory.

2.2. Objection: These successor theories are un*conceived, and thus are examples of the PUA*

A Stanfordite might claim that these historical episodes *do* instantiate the PUA. I can imagine two possible grounds for this: (a) at the earlier time, the eventual successor theories were *not* conceived in full detail. (b) The above presentation treats theories too atomistically; if instead the unit of analysis was the whole set of related hypotheses brought to bear on the phenomena, then these cases instantiate the PUA, since the whole set of successor theories was *not* conceived at the earlier time.

A Stanfordite who urges (a) points out that e.g. Ptolemy does not consider Newton's specific model of the Universe in all its detail, and therefore concludes that Ptolemy's case is part of the inductive base for the NI. A similar objection could be leveled at material caloric theorists' explicit consideration of the view that heat is motion: the scientific revolutionaries' view is less specific than the theory that would later supplant the material caloric theory, viz. the kinetic theory. Thus, the successor theory had not been truly conceived (since the predecessor lacked the successor's full detail), and therefore this case is also part of the NI's inductive base.

First, I agree that these conceived,-rejected,-then-accepted theories were often not originally conceived in the complete detail of the actual successor. However, this is not evidence that the NI can use these cases as part of its inductive base. For a realist about the successor theory would say that the previously rejected theory (e.g. 'The Earth rotates diurnally') was nonetheless true—even if the earlier version is not maximally specific. Therefore, Stanford's PUA-based argument against realism would founder, since the list of alternative hypotheses considered at the earlier time *does* contain a true (though not maximally detailed) hypothesis. Furthermore, for certain examples, this lack-of-specificity objection rests on a factual error. For example, Daniel Bernoulli *had* proposed a theory similar to the modern kinetic theory of gases in 1738 (Brush 1976, 20), a century before the kinetic theory was widely accepted.

A Stanfordite who instead appeals to (b) could stress that some of Ptolemy's arguments against the Earth's diurnal rotation use (parts of) Aristotelian physics for premises. Thus, the appropriate 'alternative' hypothesis here is not the bare claim that the Earth rotates (which she grants was conceived), but rather the conjunction of this claim

and the relevant parts of Newtonian or general relativistic dynamics. And those larger conjunctions *were* unconceived by Ptolemy, so this more holistically-conceived theory would be part of the inductive base for the NI.

This deserves two replies. First, although Ptolemy's arguments against the Earth's diurnal rotation appeal to Aristotelian physics, his other arguments do not. For example, the arguments from lack of stellar parallax and the fact that every observer sees half the zodiac are independent of Aristotelian physics. Second, more generally, this objection is in tension with Stanford's professed aversion to confirmational holism. The 'alternatives' that were unconceived in this response are not individual hypotheses, but whole conglomerations of theories. I will not weigh the pros and cons of confirmational holism here, but Stanford himself expresses anti-holist sentiments (2006, §2.2).

3. The Problem of Unconceived *Unrepresentativeness*

This section argues that many famous examples of discarded historical hypotheses resulted from projective inferences, not eliminative ones.[3] These projective inferences were often problematic because inquirers did not realize their samples were unrepresentative of the total population in relevant ways; some variable that was not previously recognized as relevant was in fact relevant. To mimic Stanford's terminology, we might call this the 'Problem of Unconceived Relevant Variables,' or the 'Problem of

[3] The view that all inductive inference actually has the form of disjunctive syllogism has been defended (Montague 1906, 281). One could argue that the inference from 'All A's observed thus far have been B' to 'All A's are B' in effect eliminates any more complex explanation of why the observed A's have been B, without explicitly listing all these more complex alternatives. Stanford himself mentions a similar possibility (2010, 234).

Unconceived Unrepresentativeness.' Stanford's PUA-based argument against realism sidelines these cases, since they do not involve eliminative inferences—even though they are *prima facie* excellent evidence for a historically based case against realism.

3.1. Examples

One example of such a case is the Galilean velocity-addition law, which was superseded by the Lorentz transformation at the beginning of the 20th Century.

The Galilean velocity-addition law is:

$$X' = X - Vt$$

(where *v* is the relative velocity between the observer and the moving object). The corresponding Lorentz Transformation is:

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Length contraction can be derived from this. The Galilean velocity-addition law (and its corollary, that the length of a rigid body is independent of its frame of reference) was presumably, for most scientists from the 17th to the end of the 19th century, seen as the conclusion of a projective argument (Newton says as much in the General Scholium, at least if being "rendered general by induction" is (a type of) projection). So it is a discarded theory that is not the result of an eliminative inference, and thus one that a Stanfordite cannot appeal to as part of the inductive base for the NI.

A second example is the classical hypothesis that the 'fixed' stars are eternal. This was widely held, presumably on projective grounds, until there was sufficient data to demonstrate that what we call 'novas' were not changes in the Earth's upper atmosphere.

The discovery of superconductivity provides a third example of this type. Before Heike Kamerlingh Onnes discovered Mercury's superconducting state in 1911, scientists projectively inferred that a body's heat capacity is proportional to its temperature, and its electrical resistivity is proportional to temperature cubed. Onnes observed that below a certain critical temperature, both quantities quickly approached zero.

Another straightforward example is the projective inference that since classical mechanics worked at many scales, it works at all scales—an inference invalidated by quantum phenomena. (This is not the claim that the laws of motion were themselves inferred projectively; rather, the claim is that the inference from 'the laws of classical mechanics hold for middle-sized dry goods and the solar system' to the conclusion 'The laws of classical mechanics will also hold for smaller spatial scales' is projective.)

3.2 Objection: These cases are evidence for the NI

An objector might suggest that these are simply paradigm examples of the PUA in action: Enlightenment scientists didn't *conceive* that the length of a moving body is inversely proportional to $\sqrt{(1 - v^2/c^2)}$. 19th C scientists didn't conceive that, for certain materials, there is some temperature below which that material's heat capacity decreases very rapidly to zero, instead of linearly. Of course, I grant that historical scientists did not conceive of these currently accepted hypotheses. However, Stanford cannot hold that these cases are part of the NI's inductive base without relinquishing some of his other commitments; in particular, he must either (a) claim that projective inferences are subject to the PUA, or (b) hold that these apparently projective inferences are actually

eliminative inferences in disguise. Stanford cannot accept either, without giving up other core claims.

Suppose a Stanfordite accepts (a), allowing that these are projective inferences, but claiming that projective inferences can be part of the NI's inductive base. I will argue that Stanford cannot do this without frustrating his desire to draw a significant epistemic difference between projective and eliminative inferences (2010; 2011; 2006, 39). For example, Stanford writes: "a consensus in favor of a given theoretical scientific belief should be regarded with considerably more suspicion when the evidence we have in support of that belief is exclusively or even just centrally abductive in character," as opposed to projective (2010, 234). (Terminological note: 'abductive evidence' constitutes the premises of an eliminative inference.) This is the core of what we could call Stanford's 'selective instrumentalism': "the *limited* skepticism thus motivated [by the NI] should certainly not extend to every scientific claim or hypothesis and may even have different force as applied to the scientific exploration of different domains" (2006, 37; my emphasis).[4] And Stanford uses this distinction when he justifies his belief that microscopic organisms exist[5] by appealing to the fact that we have "evidence of a non-eliminative character" of their existence (2006, 35; see also 33). The fact that microscopic organisms are not eliminatively inferred would not be evidence for their existence, unless we should be realists about (empirically successful, widely adopted) theoretical claims that are *projectively* inferred.[6] But if the PUA applies to projective inferences, then we should not be realists about the conclusions of projective arguments. So Stanford can

[4] See (Magnus 2010) for discussion of Stanford's piecemeal strategy.

[5] This is important to Stanford, for it distinguishes his instrumentalism from Constructive Empiricism.

only follow (a) by giving up his 'limited,' selective, or "piecemeal" instrumentalism (2006, 48). And if he does that, then his view collapses back into the classical PI (or something similar) (Magnus 2010, §3)—at least in terms of how much 'suspicion' we should cast over our current scientific theories.

Suppose a Stanfordite instead pursues (b), claiming that the above examples are actually eliminative inferences in disguise. Then there are two possibilities: either all apparently projective inferences are actually eliminative inferences, or there is something special about these apparently projective inferences that makes them eliminative inferences in disguise. If all apparently projective inferences are actually eliminative inferences, then again the distinction between projective and eliminative inference disappears, and with it Stanford's piecemeal, limited skepticism. So there must be something about these *particular* apparently projective inferences that make them eliminative. However, this appears untenable for Stanford as well. For he says that since one can "reframe" *any* projective inference as an eliminative one (2010, 234), the way to draw the line between projective and eliminative inference is by distinguishing those inferences that are "*amenable* to construal as a kind of inductive projection" from those that are not (2010, 235). That is, Stanford thinks we should view inferences that cannot be couched as projective inferences with 'considerably more suspicion' than those that

[6] Stanford makes similar remarks about the 'hypothesis of organic fossil origins': "the vulnerability of the hypothesis of organic fossil origins to any serious version of the challenge posed by the PUA has been most dramatically reduced by the fact that we have managed to supplement the fundamentally abductive sorts of evidence long available in support of it with compelling further evidence that depends instead on a more straightforward sort of inductive projection" (2010, 221).

can. This seems to rule out the 'eliminative inference in disguise' route for the Stanfordite, since all that is required for an inference to be projective in the sense Stanford considers epistemically important is that the inference *can* be framed as projective—and the above examples fit that bill.

3.3. Objection: this is merely the general problem of induction, thus not a specific *problem for* scientific *knowledge*

When discussing eliminative inference in scientific reasoning, Stanford brackets Cartesian skeptical hypotheses, even though many skeptical arguments are eliminative. His reason for doing so is that Cartesian skepticism is a problem in general epistemology, whose purported provenance is not specifically about scientific theorizing. But, he says, the arguments against scientific realism are supposed to pose a special problem for *scientific* knowledge, not for all knowledge (2006, 12-13). So, perhaps a Stanfordite would level a parallel objection against the problem of unconceived unrepresentativeness: it is just the hoary philosophical problem of induction. The problem of induction is an important problem in general epistemology, like Cartesian skepticism—but it is not a special problem for science, just as Cartesian skepticism is not a special problem for science. But (to repeat) historically-motivated anti-realism is supposed to be a special problem for science.

First reply: what's sauce for the goose is sauce for the gander. The PUA-based argument against realism may suffer from an exactly analogous problem, since disjunctive syllogism is ubiquitous: if Chrysippus is to be believed, even dogs use it. This inference form is not restricted to scientific inquiries, and having doubts about the

disjunction premise in a disjunctive syllogism is not restricted to scientific contexts, either.

Second, whether the problem of induction is a special problem for science depends on the exact meaning of 'the problem of induction.' If the problem of induction is restricted to the question 'Can we give a non-circular justification of our belief that the future will resemble the past?', then it is not a special problem for science. But the above cases of the Galilean transformation etc. do not merely extrapolate future events from past ones. Rather, they instantiate a more abstract pattern of inductive generalization: given that a limited sample of a certain kind of entity is a certain way, infer that all such entities are that way. These traits are not merely inferred to be uniform over time, but rather over a wide variety of other variables (including scale, velocity, or temperature). And this *is* a paradigmatically scientific inference: inferring from our very limited observations in the lab or field that some pattern holds for any sample of DNA, or any energetically closed system, or any two massive bodies in the universe etc., is a stereotypically scientific inference. (Of course, reasoning from a limited sample to a total population is of course not the exclusive province of science. But again, neither is disjunctive syllogism.)

Furthermore, and more importantly, the *degree* or *scale* of generalization in scientific contexts—at least for the kind of fundamental theories and mechanisms at issue in the realism debate—is typically much wider in scientific inferences than everyday inferences. Howard Stein expresses this point clearly when describing

his initial encounter with Newton's argument for universal gravitation in the *Principia*:

> "The empirical evidence available to Newton all concerned what one can reasonably describe as, first, "ordinary" behavior of "ordinary" terrestrial bodies (which of course contains no sign whatever of any such universal mutual attraction), and second, crucially, the changes of position… of eleven bright objects in the sky, and the changes in visible shape and/or luminousness … of a few of these. *To say that these are, prima facie, scanty grounds for the astoundingly far-reaching conclusion Newton came to will surely be seen as no overstatement*." (2014(?!), 2; my emphasis)
> <<This is maybe not the best quotation, because this is stressing the scantiness of the evidence, not the breadth of the conclusion; maybe break this 'commonsense vs scientific inductions' point into (a) logically stronger conclusions (b) less evidence [given the logical strength of the conclusion] for the conclusion; Stein's feeling is about (b).>>

The point is that our everyday inductive practices do not typically make conclusions about every massive body in the universe, the shape of every gene, or every species in the history of the planet. Our workaday inductions are typically to much weaker conclusions. [[There's the old logical problem that EVERY generalization is about everything: All Fs are Gs = Everything is the universe is such that either it is F and G or it's not F. But]]

In sum, however we construe 'The problem of induction,' this objection fails. If 'the problem of induction' only refers to the problem of justifying beliefs about the future on the basis of beliefs about the past, then the historical cases presented above are *not* part of the problem of induction that belongs to general epistemology. However, if 'the problem of induction' refers to any inference from a limited sample to a total population,

then such sample-based reasoning *is* central to much scientific inference—and the cases discussed in this section instantiate this type of reasoning.

4. Can't the Stanfordite accept the PI?

The previous two sections aimed to show that a proponent of the NI cannot appeal to many of the 'Greatest Hits' in the standard historically-motivated anti-realist's catalogue, without giving up other central Stanfordite commitments. A Stanfordite might respond that she can agree with everything in the previous two sections. After all, Chapters 6 and 7 of (2006) defend (something like) the PI from realist criticisms. And Stanford himself says: "I view the problem of unconceived alternatives not as competing with the traditional challenges of underdetermination and the pessimistic induction so much as bringing out what was most significant and compelling about those challenges to begin with" (2006, 45). I agree that the PUA and the NI are consistent with the PI: historically based anti-realists can make new inductions, but keep the old. However, I do not agree that the PUA captures 'what was most significant about the PI to begin with.' Furthermore, accepting the PI is inconsistent with Stanford's piecemeal instrumentalism.

4.1. Stanford over-values the PUA-based argument against realism

Stanford believes the PUA poses the most important problem for scientific realism. He writes: "the problem of unconceived alternatives… lies at the *heart* of *any* serious objection to scientific realism" (200; my emphasis). This is not an isolated moment of over-exuberance, since we find similar remarks elsewhere: "it is our vulnerability to the problem of unconceived alternatives … that is *most significant* to the debate over

scientific realism" (23; my emphasis), and "the problem of unconceived alternatives poses the *most serious* challenge to believing the claims of contemporary scientific theories" (39; my emphasis). In short, Stanford considers the PUA-based argument against realism the best argument against realism.

Let us consider Stanford's claims in turn. First, let us examine the assertion that the PUA 'lies at the heart of any serious objection to scientific realism.' If so, then if the PI is a serious objection to scientific realism, then the PUA is at the heart of the PI. I do not know precisely what Stanford means by '*X* is at the heart of any *Y*,' but presumably it entails '*X* is a necessary condition for *Y*.' So, if the PUA disappeared, then there would be no serious objections to scientific realism—including the PI. I think this is incorrect. Sections 2-3 presented classes of historical cases that are evidence for the PI but not evidence for the NI. This shows that there is some evidence against realism that is independent of the PUA. So, one might infer directly that this evidence is suffices to demonstrate that the PUA is *not* necessary for 'any serious objection to realism.' However, a Stanfordite could conceivably reply that any historically based argument against realism that omitted every historical case that did not instantiate the PUA is not 'serious' (perhaps on the grounds that it would simply leave out too many important cases from the dustbin of history). Now, I take no position on whether there are enough cases in the historical record that instantiate the problem of *conceived* alternatives, or the problem of unconceived *unrepresentativeness*, to pose a 'serious' challenge to realism—primarily because I do not wish to argue about what the threshold number for 'serious' is.

Nonetheless, this Stanfordite reply can still be answered. To say that the PUA is a necessary condition for any serious objection to realism means that if the PUA

disappeared, then all the serious objections to realism would also disappear. So let us imagine that, for historical episodes that actually involved unconceived successor theories, all the later successor theories had instead been conceived at the earlier time. For example, imagine that quantum mechanics had been explicitly formulated in 1750. Now the question is: can a serious objection to realism be posed in this hypothetical alternative history? And the answer appears to be *yes*: as Magnus (2006) and Saatsi (2009, 359) have pointed out, given the evidence available to scientists in 1750, Newtonian mechanics is better confirmed than general relativity and quantum mechanics. (No independent evidence available in 1750 suggests that the laws of motion are drastically different at scales outside what was detectable in 1750.) So scientists in 1750 (if they are rational) would have still accepted Newtonian mechanics, even if the PUA were removed. If something similar holds for many other historical cases that actually involved unconceived successor theories, then the PUA is not necessary for a serious objection to realism (unless there is *no* serious, historically based objection to realism—a route obviously unavailable to the Stanfordite).

Let us now turn to Stanford's claims that the PUA is the "most significant" or "most serious" challenge to realism. I believe it may be irresoluble whether the PUA or the PI is more important, since there may be no standard of significance or seriousness to decide the question shared between the disputants. But since Stanford makes this claim, I address it. I see at least two reasons to resist it. First, if an argument leaves out much of the *prima facie* evidence for a conclusion, and correspondingly settles for a more restricted conclusion (the moral of sections 2-3), then that argument is *prima facie* less important than a related argument that draws upon more of the relevant evidence, and

accordingly establishes a wider conclusion. Second, one could argue that 'problem of *conceived* alternatives' cases, like the geo-eccentric hypothesis, are a more serious problem for realism than cases that instantiate the PUA. Why? Cases like the Earth's diurnal rotation and Daniel Bernoulli's kinetic theory of gases show that *even when* we explicitly consider the correct theory—i.e. the truth is 'staring us in the face'—we can *still* reject it. In my opinion, that makes us look even more epistemically inept than an inability to conceive of the true theory, when the true theory is conceptually distant from our current list of live options. To be clear, I do not think these two reasons conclusively establish that the PI is definitely more significant than the PUA (whatever 'significance' comes to). However, they do present a challenge that a Stanfordite must answer, if she wishes to maintain that the PUA presents the most significant challenge to realism.

4.2. Accepting the PI is inconsistent with Stanford's piecemeal instrumentalism

The argument for the claim in above sub-heading is straightforward, since all the pieces were presented in 3.1 above. Stanford wants his criticism of realism to be selective or piecemeal. This is achieved by casting "considerably more suspicion" on widely accepted, empirically successful fundamental theories that are eliminatively inferred than those that are projectively inferred. Yet §3 showed that many theories that are evidence for the PI's inductive base were projectively inferred. Thus, anyone who accepts the PI as a persuasive argument against realism cannot (in the absence of special pleading) simultaneously endorse Stanford's brand of piecemeal skepticism about scientific theories.

5. Conclusion

A historically motivated anti-realist who only appeals to the PUA and NI needlessly limits both the evidence for her view, and the scope of her anti-realist conclusion, since she omits all historical cases where either the successors were conceived, or the theories were the result of projective inferences. The most natural attempts to incorporate such cases into the PUA-based argument against realism fall afoul of Stanford's other commitments.

References

Barber, Bernard (1961). "Resistance by Scientists to Scientific Discovery," *Science* 134:596-601

Black, Joseph (1806). *Lectures on the Elements of Chemistry*, Vol. I. Philadelphia: Mathew Carey.

Brush, Steven (1976). *The Kind of Motion We Call Heat*, Book 1, New York: New Holland Press.

Hook, Ernest (ed.) (2002). *Prematurity in Scientific Discovery*. Berkeley: University of California Press.

Linnaeus, Carolus (1964 [1735]). *Systema Naturae [General System of Nature*]. Facsimile of the first edition with an introduction and an English translation of the observations by M. S. J. Engel Ledeboer and H. Engel (Nieuwkoop, Holland: B. de Graff, 1964).

Magnus, P. D. (2006). "What's New About the New Induction?" *Synthese* **148**: 295-301.

Magnus, P. D. (2010). "Inductions, Red Herrings, and the Best Explanation for the Mixed Record of Science," *British Journal for the Philosophy of Science* **61**: 803-819.

Metcalfe, Samuel (1859). *Caloric: Its Mechanical, Chemical, and Vital Agencies in the Phenomena of Nature*, Vol. I. Philadelphia: J.B. Lippincott.

Saatsi, Juha (2009), "Grasping at Realist Straws," *Metascience* **18**: 355-363.

Stanford, P. Kyle (2006). *Exceeding Our Grasp*. New York: Oxford University Press.

Stanford, P. Kyle (2010). "Getting Real: The Hypothesis of Organic Fossil Origins," *The Modern Schoolman* **87**: 218-243.

Stanford, P. Kyle (2011). "Damn the Consequences: Projective Evidence and the Heterogeneity of Scientific Confirmation," *Philosophy of Science* **78**: 887-899.

Why I Am Not a Methodological Likelihoodist*

By Greg Gandenberger

March 14, 2014

Abstract

Methodological likelihoodism is the view that it is possible to provide an adequate self-contained methodology for science on the basis of likelihood functions alone. I argue that methodological likelihoodism is false by arguing that an adequate self-contained methodology for science provides good norms of commitment vis-à-vis hypotheses, articulating minimal requirements for a norm of this kind, and proving that no purely likelihood-based norm satisfies those requirements.

Introduction

One of the guiding ideas in the philosophy of induction is that "saving the phenomena is a mark of truth" (Norton, 2005, 11). In other words, a hypothesis is confirmed to the extent that it correctly predicts what is observed; as Milne puts it, "prediction and confirmation are two sides of the same coin" (1996, 23).

Likelihoodism provides principles of evidential relevance and evidential favoring that accord with this idea. Its primary principle of evidential favoring is the *Likelihood Principle*, which says that the evidential meaning of a datum

*Thanks to Jake Chandler, Branden Fitelson, Clark Glymour, Satish Iyengar, Michael Lew, Edouard Machery, John Norton, Teddy Seidenfeld, and Jim Woodward for discussions that contributed to the development of this paper.

E for a set of hypotheses H depends only on how well each of those hypotheses predicts that datum—more precisely, on the *likelihood function* $\Pr(E|H)$[1] considered as a function of $H \in \mathbf{H}$, up to a constant of proportionality. Its primary principle of evidential favoring is the *Law of Likelihood*, according to which E favors H_1 over H_2 when and to the degree that the *log-likelihood ratio* $\mathcal{L} = \log[Pr(E|H_1)/Pr(E|H_2)]$ is greater than zero.[2]

Methodological likelihoodists such as Edwards (1972), Royall (1997), and Sober (2008) go beyond simply accepting the Likelihood Principle and the Law of Likelihood: they claim that it is possible to provide an adequate self-contained[3] methodology for science on the basis of likelihood functions alone. They aim to provide a methodology that combines the main advantages of Bayesian and frequentist methodologies without their respective disadvantages. Like Bayesian and unlike frequentist methods, likelihoodist methods conform to the Likelihood Principle, for which there are strong arguments (Gandenberger, forthcoming). Like frequentist and unlike Bayesian methods, likelihoodist methods avoid appeals to prior probabilities, which are often contentious.

[1] A more subtle account is needed when the sample space is continuous, so that $\Pr(E|H) = 0$ for a typical datum; see (Hacking, 1965, 57, 66–70), (Berger and Wolpert, 1988, 32–6), and (Pawitan, 2001, 23–4). In principle, this complication can be ignored in the context of any real experiment: real measurement techniques have finite precision, so real sample spaces are always discrete.

[2] The Law of Likelihood is often stated in terms of likelihood ratios rather than log-likelihood ratios. Nothing substantive hangs on this difference. Strictly speaking, the Law of Likelihood should be understood as the claim that evidential favoring increases monotonically with the likelihood ratio. Different monotonic functions of the likelihood ratio produce different permissible measurement scales. I use a logarithmic scale for ease of interpretation: zero indicates evidential neutrality; positive values indicate that the evidence favors H_1 over H_2 while negative values indicate the opposite; and the degree of evidential favoring provided by a pair of independent data is simply the sum of the degrees of evidential favoring provided by each datum individually. The base of the logarithms is immaterial.

[3] The claim that this methodology is self-contained is not meant to exclude methodological pluralism a la (Sober, 2008, 3, 356–8). Methodological likelihoodists need not believe that methods based on likelihood functions alone are appropriate for *all* scientific problems. However, they must believe that they are appropriate for *some* scientific problems, and not merely in the sense that they are appropriate when they would give the same answer as a reasonable Bayesian or frequentist method or in the sense that their outputs are useful as inputs for some other method, such as Bayesian updating. A common pluralist view that qualifies as a form of methodological likelihoodism is that Bayesian methods are appropriate when prior probabilities are "available" in some sense, while likelihoodist methods are appropriate when they are not, and that they are often unavailable in science (see e.g. Sober 2008, 32).

The purpose of this paper is to argue that methodological likelihoodism should nevertheless be rejected.

My argument against methodological likelihoodism rests on the following premises.

Premises

(1) An adequate self-contained methodology for science provides good norms of commitment vis-à-vis hypotheses.

(2) If there are good norms of commitment based on likelihood functions alone, then some rule of the following form is among them, where T is your total relevant evidence:[4]

> **P**roportion **R**elative **A**cceptance to (a **F**unction of) the **E**vidence (PRAFE): Accept H_1 over H_2 to the degree $f(\mathcal{L}) = f(\log[\Pr(T|H_1)/\Pr(T|H_2)])$, where f is some nondecreasing function such that $f(0) = 0$ and $f(a) > 0$ for some a.

(3) A good norm of commitment vis-à-vis hypotheses is compatible with the following rules:

(3A) Do not prefer H_1 to H_2 and H_3 to H_4 if H_1 is logically equivalent to H_4 and H_2 is logically equivalent to H_3 (where preferring one hypothesis to another is equivalent to accepting the former over the latter to a positive degree).

(3B) Accept (H_1 or H_2) over H_3 to a degree greater than that to which you accept H_1 over H_3 when the following conditions are met:

i. H_1 and H_2 are mutually exclusive,

[4]The notion of relevant evidence can be formalized in terms of sufficient statistics (see Halmos and Savage 1949).

ii. the degree to which you accept H_1 over H_3 is well-defined, and

iii. the degree to which you accept H_2 over H_3 is well-defined and is not $-\infty$.

(4) A good norm of commitment vis-à-vis hypotheses is compatible with the following rule:

(4A) Do not prefer H_1 to $\sim H_1$ and $\sim H_2$ to H_2 if H_1 and H_2 are logically equivalent given your total evidence.

(3) and (4) each entails that no rule of the form given by (PRAFE) is a good norm of commitment vis-à-vis hypotheses. Thus, the conjunction of (1), (2), and either (3) or (4) entails that methodological likelihoodism is false.

I argue for (1)–(4) in Sections 1–4, respectively. I prove that (3) and (4) entail that no rule of the form given by (PRAFE) is a good norm of commitment vis-à-vis hypotheses in Appendices A and B, respectively. In Section 5 I respond to attempts to defend likelihoodist methods with reliabilist arguments.

1 Premise (1): An adequate self-contained methodology for science provides good norms of commitment vis-à-vis hypotheses

Science should help guide our commitments vis-à-vis hypotheses. It is not enough to say something about how the data are related to the hypotheses; we need to be able to "detach" the evidence and say something about the hypotheses themselves in light of the data. It would be odd for the author of a scientific paper to say that he or she does not care about evaluating the hypotheses he or she considers, but only wishes to assess how the data bear on

them as evidence.[5]

Some methodological likelihoodists, such as Edwards,[6] at least suggest that purely likelihood-based methods can be used to guide our commitments vis-à-vis hypotheses. Others, such as Royal and Sober, are careful not to claim more for purely likelihood-based methods than that they correctly characterize data as evidence (e.g. Royall 1997, 3; Sober 2008, 32). But what are we to do with a characterization of data as evidence if not to use it to guide our commitments? And how are we to use it to guide our commitments without appealing to information not given by the likelihood function?

Royall and Sober provide no explicit answers to these questions. Royall takes the value of a correct characterization of data as evidence for granted.[7] Sober does not take it for granted, but he does not provide a clear argument for it either. He says that it is not enough to show that the Law of Likelihood "conforms to, and renders precise and systematic, our use of the informal concept" of evidential favoring: "what matters," he writes, "is whether [the Law of Likelihood] isolates an epistemologically important concept" (Sober, 2008, 35). I agree that it is not enough to vindicate methodological likelihoodism to show that the Law of Likelihood captures our informal concept of evidential favoring. But, depending on one's views about epistemological importance, it may also not be enough to show that the Law of Likelihood isolates an epistemologically important concept. The Law of Likelihood could be epistemologically important because, for instance, it is useful for explaining the so-called conjunction

[5]Of course, not every scientific paper should include an evaluation of the hypotheses considered therein. There may be relevant data from other sources, and the author of the paper may wish to leave the evaluation to his or her readers. These points are compatible with my claim that hypothesis evaluation is ultimately indispensable.

[6]Edwards states the Law of Likelihood in terms of support (1972, 30), but he describes it as providing the basis for a system of inference (e.g. 1972, 7) and relative degrees of belief (e.g. 1972, 28) without explanation or argument.

[7]Royall begins his (1997) with the bare assertion that the "most important task" of statistics "is to provide objective quantitative alternatives to personal judgement for interpreting the evidence produced by experiments and observational studies" (xi). I have found no argument for this claim in his writings.

fallacy.[8] It would not follow that the Law of Likelihood provides useful guidance for our commitments vis-à-vis hypotheses, as methodological likelihoodism requires.

Sober seems to take himself to show that the Law of Likelihood does isolate an epistemologically important concept. However, he does not explain how he takes himself to do so. It is a plausible guess that he takes himself to do so in his use of the Law of Likelihood to address seemingly well-formed and important question such as the following (Sober, 2008, 107–8):

- Are the imperfect adaptations that organisms exhibit evidence that they were not produced by an intelligent designer?
- Is the fact that bears in cold climates have longer fur than bears in warm climates evidence that fur length evolved by natural selection as an adaptive response to ambient temperature?
- Are the similarities that species exhibit evidence that they stem from a common ancestor?

The Law of Likelihood can indeed be used to provide defensible answers to these questions (see Gandenberger forthcoming, Gandenberger unpublished). However, the following questions remain: what are we to do with answers to these questions if not to use them to guide our commitments vis-à-vis the relevant hypotheses? And how are we to use them to provide such guidance without appealing to information not given by the likelihood function?

[8] As an example of the conjunction fallacy, most people give a higher probability to the statement that a character named Linda is a feminist bank teller then to the proposition that Linda is a bank teller. Because the population of feminist bank tellers is necessarily a subset of the population of bank tellers, these judgments are probabilistically incoherent. One possible explanation for this phenomenon is that people are responding to the fact that the vignette told about Linda favors the statement that she is a feminist bank teller over its negation (or confirms it) more than it favors the statement that she is a bank teller over its negation. See Tentori et al. (2013) for empirical evidence that seems to support this explanation.

Methodological likelihoodists have two options: (1) claim that science need not provide guidance for our commitments vis-à-vis hypotheses, or (2) provide good norms of commitment vis-à-vis hypotheses that are based on likelihood functions alone. Option (1) flies in the face of the commonsense idea that we do science in order to learn about the world (at least in some attenuated sense)[9] or to improve our ability to predict and control some part of it. It would be difficult to justify allocating time and tax dollars to science if all it could do were to generate data and hypotheses and tell us how that data is related to those hypotheses as evidence, without thereby giving us any guidance about what to believe or do. Traditionally, philosophers of science have sought a theory of evidence or confirmation so that they could use that theory to evaluate hypotheses in a principled way. The idea that characterizations of data as evidence are valuable in themselves is an unfortunate byproduct of this pursuit.

I take up option (2) in the next section.

2 Premise (2): If there are good purely likelihood-based norms of commitment, then (PRAFE) is among them

I argued in the previous section that making good on methodological likelihoodism requires providing a good purely likelihood-based norm of commitment.

A reasonable starting point for an attempt to provide such a norm is Hume's dictum that a wise person proportions his or her belief to his or her (total) evidence (1825, 111, paraphrased). We can increase the plausibility of this

[9]It is compatible with my claim in the section, for instance, that we do science only to learn approximate truths about observable phenomena, as some scientific anti-realists claim (e.g. Van Fraassen 1980).

already plausible dictum by generalizing it in two ways. First we can replace "belief" with "acceptance." The word "acceptance" here could be understood in a purely doxastic way, as indicating "what one holds in one's head," so to speak. Alternatively, it could be understood in a "behavioristic" way, as indicating something about what one would do in certain circumstances. I aim to show that my generalization of Hume's dictum has disastrous consequences on either interpretation, thereby casting doubt on the possibility of a good purely likelihood-based norm of commitment of either the doxastic or the behavioristic kind. I assume only that degrees of acceptance correspond to a qualitative preference ordering in the following way: accepting H_1 over H_2 to a positive degree indicates that one prefers H_1 to H_2, doing so to degree zero indicates that one has no preference between H_1 and H_2, and doing so to a negative degree indicates that one prefers H_2 to H_1.

We can generalize Hume's dictum in a second way by saying merely that a wise person proportions his or her acceptance to some function of the evidence, where that function satisfies the following mild constraints:

- it is **nondecreasing**, so that an increase in absolute value for evidential favoring without a change in sign[10] never leads to a decrease in degree of acceptance;
- it is **calibrated** in the sense that neutral evidence ($\log[\Pr(T|H_1)/\Pr(T|H_2) = 0$, i.e. $\Pr(T|H_1) = \Pr(T|H_2)$) leads to neutrality of acceptance (neither preferring H_1 to H_2 nor vice versa);
- it is **nontrivial** in the sense that it would lead one to accept H_1 over H_2 given sufficiently strong evidence favoring the former over the latter.

[10]The phrase "without a change in sign" is necessary because I do not assume that the function f is symmetric about zero. One might wish to add this assumption, but I do not need it.

Generalizing Hume's dictum in these two ways leads to the following class of purely likelihood-based norms of commitment vis-à-vis hypotheses, where T is one's total relevant evidence.

> **P**roportion **R**elative **A**cceptance to (a **F**unction of) the **E**vidence (PRAFE): Accept H_1 over H_2 to the degree $f(\mathcal{L}) = f(\log[\Pr(T|H_1)/\Pr(T|H_2)])$, where f is some nondecreasing function such that $f(0) = 0$ and $f(a) > 0$ for some a.

I cannot think of a more plausible yet nontrivial way to map the degrees of evidential favoring that the Law of Likelihood provides onto real-valued degrees of relative acceptance. There are no rival proposals in the literature to consider because methodological likelihoodists either deny that their methods provide guidance for belief or action (e.g. Royall and Sober) or suggest that they do provide such guidance but fail to provide a definite account (e.g. Edwards). If methodological likelihoodists wish to claim that there are good purely likelihood-based norms of commitment of a different form, then they need to say explicitly what those norms are and how they are good. In the meantime, (PRAFE)'s generality and intuitive plausibility warrant the claim that if there are good purely likelihood-based norms of commitment, then they include some norm of the form it provides.

Methodological likelihoodists may not be able to escape the difficulties for (PRAFE) that I present below even if there are purely likelihood-based norms of commitment more plausible than (PRAFE). I have argued for a particular kind of norm only because some constraints are necessary for proving definite results. But the problematic results for (PRAFE) that I present do not seem to depend on any quirk of (PRAFE) that could easily be removed, but rather from the fact that likelihood functions (unlike probability distributions) respect neither entailment relations among hypotheses nor logical equivalence among

hypotheses given one's evidence. For that reason, it seems likely that any purely likelihood-based norm of commitment would suffer from similar problems.

3 Premise (3): A good norm of commitment is compatible with (3A) and (3B)

I have argued that an adequate self-contained methodology for science provides good norms of commitment vis-à-vis hypotheses and that if there are any purely likelihood-based norms of this kind then a norm of the form given by (PRAFE) is among them. It follows that methodological likelihoodism is false if no norm of the form given by (PRAFE) is a good one. In this section I argue that no norm of the form given by (PRAFE) is a good one because any norm of this kind can force one to violate either (3A) or (3B):

(3A) Do not prefer H_1 to H_2 and H_3 to H_4 if H_1 is logically equivalent to H_4 and H_2 is logically equivalent to H_3 (where preferring one hypothesis to another means accepting the former over the latter to a positive degree).

(3B) Accept (H_1 or H_2) over H_3 to a degree greater than that to which you accept H_1 over H_3 when the following conditions are met:

i. H_1 and H_2 are mutually exclusive,

ii. the degree to which you accept H_1 over H_3 is well-defined, and

iii. the degree to which you accept H_2 over H_3 is well-defined and is not $-\infty$.

These rules are compelling. Take (3A). This rule seems innocuous in applications. For instance, it says not to prefer "all ravens are black" to "some ravens are white" while at the same time preferring "some white things are ravens"

to "all non-black things are non-ravens." (Note that "some white things are ravens" is logically equivalent to "some ravens are white" and "all non-black things are non-ravens" is logically equivalent to "all ravens are black.") After all, one's preference between a pair of propositions should not depend on the form in which those propositions are stated.

Moreover, (3A) is completely trivial under various possible formalization of the notion of relative acceptance. For instance, if we interpret the degree to which one accepts A over B as one's log-odds $\log[\Pr(A)/\Pr(B)]$, then (3A) follows from the fact that probabilities do not change under substitution of logical equivalents.[11] In fact, (3A) follows from any formalization that allows substitution of logical equivalents.

Now take (3B). Roughly speaking, this rule directs one to accept a disjunction over an alternative claim more than one accepts one of its disjuncts over that claim, provided that one is not willing to dismiss the other disjunct completely relative to that claim. The restriction to cases in which the degree to which one accepts H_2 over H_3 is not $-\infty$ rules out cases in which one completely rejects H_2 relative to H_3. Again, this rule seems innocuous in applications. For instance, it directs one to accept "either all ravens are black or some white and the rest of black" over "some ravens are red" to a degree greater than that to which one accepts "all ravens are black" over "some ravens are red," provided that the degree to which one accepts "all ravens are black" over "some ravens are red" is well-defined and the degree to which one accepts "some ravens are white and the rest are black" over "some ravens are red" is well-defined and

[11] It is arguably permissible in some sense for subjective degrees of belief to vary under substitution of logical equivalents. For instance, one would hardly blame a person of average mathematical ability who was attempting to assess the size of a cubic box for assigning different probabilities to the proposition that each side of the box is 27 inches long and the proposition that the box has volume $19,683$ in.^{3} (Rescorla, unpublished, 18–9), even though those hypotheses are equivalent. But this case does involve a failure to be fully rational; it is just a failure of logical omniscience rather than a failure of probability assessment. Thus, though excusable, it is not rationally permissible in any strong sense.

is not $-\infty$. This application of (3B0 seems obligatory. After all, "either all ravens are black or some are white and the rest are black" encompasses more possibilities than "all ravens are black," so it makes sense to accept the former over some third claim to a greater degree than letter, provided that one gives any credence at all to the additional possibilities it encompasses.

Like (3A), (3B) would hold under a variety of possible formalizations of the notion of relative acceptance. For instance, if we again interpret the degree to which one accepts A over B as one's log- odds $\log[\Pr(A)/\Pr(B)]$, then (3B) follows from the axioms of probability. Probabilities obey finite additivity, meaning that $\Pr(H_1 \text{ or } H_2) = \Pr(H_1) + \Pr(H_2)$ when H_1 and H_2 are mutually exclusive. It follows that $\log[\Pr(H_1 \text{ or } H_2)/\Pr(H_3)] > \log[\Pr(H_1)/\Pr(H_3)]$ when $\Pr(H_2)/\Pr(H_3) > 0$ and H_1 and H_2 are mutually exclusive. An analogous argument would work under any analogous interpretation that uses an additive (or superadditive)[12] calculus.

It is possible to give a very simple argument that (PRAFE) forces one to violate (3B) without assuming (3A). However, this argument makes an objectionable assumption. Suppose you were to run a completely uninformative experiment: you have me flip a coin with unknown bias p for heads but to report "heads" regardless of how it lands. Then $\Pr(E|p = p^*) = 1$ for all $0 \leq p^* \leq 1$. The Law of Likelihood entails that that outcome is neutral between any pair of $H_1 : 0 \leq p < 1/3$, $H_2 : 1/3 \leq p < 2/3$, and $H_3 : 2/3 \leq p < 1$. But it also implies that it is neutral between (H_1 or H_2) and H_3. This combination of claims is not problematic as long as we are only talking about evidential favoring. But using (PRAFE) to translate talk about evidential favoring into talk of acceptance yields violations of (3B).

[12] A superadditive calculus f such as the Dempster-Shafer calculus (Dempster, 1968)) is one whose axioms guarantee only that $f(H_1 \text{ or } H_2) \geq f(H_1) + f(H_2)$ when H_1 and H_2 are mutually exclusive.

If you find this argument against (PRAFE) convincing, then so much the better for my thesis. I do not place much weight on it because it has a weak point. One could claim that the size of the interval for p that a hypothesis posits is relevant to its assessment, either as part of the total relevant evidence with respect to that hypothesis or as a factor apart from the evidence that should play some role in a rule such as (PRAFE) for translating degrees of favoring into degrees of acceptance. I suspect that this claim is unsustainable, but I would rather make it moot than argue against it. I do so by providing a more elaborate argument in which I apply (PRAFE) only to intervals of equal sizes in two different parameterizations of the hypothesis space and use (3B) to generate a violation of (3A).

That argument is given in Appendix A. Here is roughly how it goes. I construct a hypothetical experiment the outcome of which is evidentially neutral between hypotheses A, B, and C according to the Law of Likelihood. By stipulation, that outcome is one's only evidence about those hypotheses. (PRAFE) thus requires you to be neutral between hypotheses A, B, and C. (3B) thus requires you to prefer (A or B) over C. I then consider an alternative set of hypotheses A', B', and C' between which the outcome is also evidentially neutral according to the Law of Likelihood. By an analogous argument, (PRAFE) and (3B) require you to prefer (A' or B') over C'. It thereby requires you to violate (3A). For the hypotheses are constructed so that (A or B) is logically equivalent to C' and (A' or B') is logically equivalent to C. Thus, (PRAFE) forces you to violate either (3A) or (3B).

Because (PRAFE) can force[13] you to violate either (3A) or (3B), the fact that (3A) and (3B) are compelling warrants the claim that (PRAFE) is not a good norm of commitment vis-à-vis hypotheses. It follows from this claim together

[13] "(PRAFE) can force you to violate..." should be understood as shorthand for "any norm of the form given by (PRAFE) can force you to violate..."

with the claims I argued for in the previous two sections that methodological likelihood is false.

4 Premise (4): A good norm of commitment is compatible with (4A)

In this section I argue that (PRAFE) is not a good norm of commitment for a second reason: it can force one to violate (4A):

> (4A) Do not prefer H_1 to $\sim H_1$ and $\sim H_2$ to H_2 if H_1 and H_2 are logically equivalent given your total evidence.

(4A) is compelling. It is similar to (3A) in that it requires degrees of acceptance to respect a certain kind of logical equivalence. Is stronger than (3A) in that requires only logical equivalence given one's evidence. It is weaker than (3A) in that it has implications only for preferences between hypotheses and their negations.

Like (3A) and (3B), (4A) seems innocuous in applications. It prohibits someone who knows that no ravens are red from preferring "all ravens are either black or red" to its negation while dispreferring "all ravens are black" to its negation. For someone who knows that no ravens are red, these hypotheses have the same content and thus should be assessed alike.

Like (3A), (4A) is completely trivial under any formalization of relative acceptance that allows substitution of logical equivalents.

I prove that (PRAFE) can force one to violate (4A) in Appendix B. Here is roughly how the proof goes. Let H_1 be the conjunction of some proposition A with E, and let H_2 be just the proposition A. Suppose that E is one's total relevant evidence. I show that for any constant a, one can construct a probability distribution over A and E such that the log-likelihood ratios of H_1

against $\sim H_1$ and $\sim H_2$ against H_2 both exceed a. Thus, given any norm of the form given by (PRAFE) there is a possible experimental outcome that would lead one to prefer H_1 over $\sim H_1$ and $\sim H_2$ over H_2. In this way, (PRAFE) can force you to violate (4A).

Because (PRAFE) can force you to violate (4A), the fact that (4A) is compelling warrants the claim that (PRAFE) is not a good norm of commitment vis-à-vis hypotheses. It follows from this claim together with the claims I argued for in Sections 1 and 2 that methodological likelihoodism is false.

It is worth noting that one could avoid violating (4A) by adopting a restricted version of (PRAFE) that applies only to statistical hypotheses—that is, hypotheses that are simply about the stochastic properties of the data-generating mechanism. H_1 in my proof—the conjunction of some proposition A with E itself—is not a statistical hypothesis because it makes a direct statement about the outcome produced by the data-generating mechanism. Some likelihoodists do restrict the Law of Likelihood in this way (e.g. Hacking 1965, 59 and Edwards 1972, 57). However, they seem to do so not for any principled reason but simply because they have statistical applications in mind. It is not clear that the restriction has any principled basis. Moreover, it has the unfortunate consequence of restricting the scope of the Law of Likelihood substantially. For instance, it would not allow one to apply the Law of Likelihood to high-level, substantive scientific theories, as Sober does with the theory of evolution and the theory of intelligent design (Sober, 2008). In addition, it does not address the fact that (PRAFE) can force one to violate either (3A) or (3B). Thus, restricting the Law of Likelihood to statistical hypotheses has a high cost and is insufficient to avoid the major difficulties for methodological likelihoodism presented in this paper.

5 Against a reliabilist response

The fact that (PRAFE) can force one to violate either (3A) or (3B) and (4A) disqualifies it from consideration as a general principle of rationality. One might attempt to rescue methodological likelihoodism by lowering one's standards. Perhaps no purely likelihood-based norms of commitment are among the canons of rationality, but such norms are nevertheless useful in practice when deployed judiciously. This move may not appeal to most philosophers, but similar moves are common among statisticians (e.g. Chatfield 2002, Kass 2011, and Gelman 2011).

The idea that (PRAFE) is useful when deployed judiciously is plausible only if it has some redeeming quality that at least partially compensates for the fact that it is inconsistent with the conjunction of (3A) and (3B) and with (4A). What could that redeeming quality be? Here are four candidates from Royall (2000, 760):

(I) intuitive plausibility,

(II) consistency with other axioms and principles,

(III) objectivity, and

(IV) desirable operational implications.

I am willing to grant (I)-(III) for the sake of argument. Those virtues are not sufficient, even jointly, to vindicate methodological likelihoodism. It still needs to be shown that methods based on likelihood functions alone can provide useful guidance for our commitments vis-à-vis hypotheses.

(IV) is prima facie more promising. It refers to the purported fact that purely likelihood-based methods are guaranteed to perform well in certain senses in the indefinite long run if used over and over again with varying data. Appeals

to guarantees about long-run performance are the hallmark of frequentism, but Bayesians cite such results as well, perhaps most often in the form of convergence theorems (e.g. Doob 1949). The exact significance of various facts about long-run operating characteristics is a matter of dispute, but there is no disputing the basic idea that we want techniques that we can reasonably expect to yield good results.

Unfortunately for this line of response, likelihoodist appeals to (IV) generate many problems. The most damning of these problems is that *the operating characteristics that likelihoodists appeal to are not operating characteristics of purely likelihood-based methods.* Instead, they are operating characteristics of methods that use likelihood functions in a frequentist way.

Let me explain. By definition, purely likelihood-based methods are not sensitive to differences between experimental outcomes that are not reflected in the likelihood function. One such fact concerns the distinction between *fixed* and *random* hypotheses. Fixed hypotheses are specified without reference to the data, while random hypotheses are specified in terms of the data. For instance, the hypothesis that the mean of the distribution that produced the data is *zero* is a fixed hypothesis, while the hypothesis that it is the *sample average* (the sum of the data values divided by the number of data values) is a random hypothesis, because the value of the sample average depends on the data while the value of the number zero does not.

By contrast, frequentist methods violate the likelihood principle by being sensitive to the distinction between fixed and random hypotheses. A frequentist may draw different conclusions about the hypothesis that the mean of a distribution is zero depending on whether he or she set out to test the hypothesis that the mean is zero or set out to test the hypothesis that the mean is the sample average, which turned out to be zero.

Whether sensitivity to the distinction during fixed and random hypotheses is a good feature for a method to have or not is a topic for another occasion. The key points for present purposes are (1) such sensitivity cannot be present inpurely likelihood-based methods, and (2) it is necessary for the long-run operating characteristics that likelihoodists erroneously cite in support of their methods. I will illustrate these claims for the *universal bound*, which is the fact that the probability of a likelihood ratio of at least k for any given fixed, false hypothesis against the true hypothesis is at most $1/k$ (i.e., $\Pr_{H_0}(\Pr(E|H_1)/\Pr(E|H_0) \geq k) < 1/k$) (Royall, 2000, 762-3). The same point holds for other results can concerning the performance characteristics of methods based on likelihood functions, including both the tighter bounds that Royall derives for specific distributions (2000) and likelihood ratio convergence theorems (Hawthorne, 2012).

An example[14] due to (Armitage, 1961) is a counterexample to a generalized version of the universal bound that applies to fixed as well as random hypotheses. I will simply describe the main features of the Armitage example here; see (Cox and Hinkley, 1974, 50–1) for details. The example involves taking observations until the sample average $\bar{x}$ is at least a specified distance away from zero. That distance decreases as the number of observations increases. It does so at a rate that is fast enough that the experiment is guaranteed to end in finite time,[15] but slow enough to ensure that according to the Law of Likelihood its final outcome strongly favors the hypothesis that the true mean equals x-bar over the hypothesis that it equals zero. For any k, there is an experiment of this kind such that $\Pr_{H_0}(\Pr(E|H_1)/\Pr(E|H_0) \geq k)$ is 1—a maximally severe violation of the universal bound. I have argued elsewhere (Gandenberger, unpublished) that this example should not be regarded as a counterexample to the Law of Likelihood

[14]Strictly speaking, Armitage provides a class of examples rather than a single example. I am using the word "example" as a convenient shorthand.

[15]Technically, the experiment ends "almost surely" (i.e., with probability one) in finite time.

itself. However, it is a counterexample to attempts to use the universal bound to support the use of purely likelihood-based methods.

Likelihood functions do not distinguish between fixed and random hypotheses, so purely likelihood-based methods cannot distinguish between them either. Thus, results such as the universal bound that hold only for fixed hypotheses do not support the use of purely likelihood-based methods. Methodological likelihoodists who wish to claim that purely likelihood-based methods are useful when deployed judiciously need to find some other support for that view.

6 Conclusion

Methodological likelihoodism is true only if (PRAFE) is a good purely likelihood-based norm of commitment. (PRAFE) is not a good purely likelihood-based norm of commitment because it can force one to violate both the combination of (3A) and (3B) and (4A). Therefore, methodological likelihoodism is false. The results concerning long-run operating characteristics that methodological likelihoodists sometimes cite in support of their methods do not help their cause because those results concern frequentist methods that use likelihood functions rather than purely likelihood-based methods.

References

Armitage, Peter. 1961. "Contribution to 'Consistency in Statistical Inference and Decision'." *Journal of the Royal Statistical Society. Series B (Methodological)* 23:30–1. ISSN 00359246.

Berger, James and Wolpert, Robert. 1988. *The Likelihood Principle*, volume 6 of *Lecture Notes—Monograph Series*. Beachwood, OH: Institute of Mathematical Statistics, 2nd edition.

Chatfield, Chris. 2002. "Confessions of a pragmatic statistician." *Journal of the Royal Statistical Society: Series D (The Statistician)* 51:1–20. ISSN 1467-9884. doi:10.1111/1467-9884.00294.

Cox, David and Hinkley, David. 1974. *Theoretical Statistics*. London: Chapman and Hall.

Dempster, AP. 1968. "A generalization of Bayesian inference (with discussion)." *Journal of the Royal Statistical Society Series B. v30 i2* 205–247.

Doob, Joseph L. 1949. "Application of the theory of martingales." *Le calcul des probabilites et ses applications* 23–27.

Edwards, A.W.F. 1972. *Likelihood. An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference. [Mit Fig. U. Tab.].* Cambridge University Press.

Gandenberger, Greg. forthcoming. "A New Proof of the Likelihood Principle." *The British Journal for the Philosophy of Science* .

—. unpublished. "New Responses to Three Counterexamples to the Likelihood Principle."

Gelman, Andrew. 2011. "Bayesian Statistical Pragmatism." *Statistical Science* 26:10–11. doi:10.1214/11-STS337C.

Hacking, I. 1965. *Logic of Statistical Inference*. Cambridge University Press. ISBN 9780521290593.

Halmos, Paul R. and Savage, L. J. 1949. "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics." *The Annals of Mathematical Statistics* 20:225–241. doi:10.1214/aoms/1177730032.

Hawthorne, James. 2012. "Inductive Logic." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

Hume, David. 1825. *Essays and Treatises on Several Subjects*, volume 2. Edinburgh: Bell and Bradfute.

Kass, Robert E. 2011. "Statistical inference: The big picture." *Statistical science: a review journal of the Institute of Mathematical Statistics* 26:1.

Lewis, David. 1981. "A subjectivists guide to objective chance." In *Ifs*, 267–297. Springer.

Milne, Peter. 1996. "log[P(h/eb)/P(h/b)] Is the One True Measure of Confirmation." *Philosophy of Science* 63:pp. 21–26. ISSN 00318248.

Norton, John. 2005. "A Little Survey of Induction." In Peter Achinstein (ed.), *Scientific Evidence: Philosophical Theories and Applications*, 9–34. Johns Hopkins University Press.

—. 2008. "Ignorance and Indifference." *Philosophy of Science* 75:45–68. doi: 10.1086/587822.

Pawitan, Y. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford science publications. OUP Oxford. ISBN 9780198507659.

Rescorla, Michael. Unpublished. "Some Epistemological Ramifications of the Borel-Kolmogorov Paradox." .

Royall, Richard. 2000. "On the Probability of Observing Misleading Statistical Evidence." *Journal of the American Statistical Association* 95:pp. 760–768. ISSN 01621459.

Royall, R.M. 1997. *Statistical Evidence: A Likelihood Paradigm*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Sober, Elliott. 2008. *Evidence and Evolution: The Logic Behind the Science.* Cambridge University Press.

Tentori, K, Crupi, V, and Russo, S. 2013. "On the determinants of the conjunction fallacy: Confirmation versus probability." .

Van Fraassen, B.C. 1980. *The Scientific Image.* Clarendon Library of Logic and Philosophy. Clarendon Press. ISBN 9780198244271.

A Proof that (PRAFE) can force one to violate either (3A) or (3B)

Proof

Suppose a mad genius has mixed water and wine in a bottle. You know only that the ratio r of water to wine is in the interval $(1/2, 2]$. The mad genius knows the value of r but refuses to tell it to you. He does agree to perform three independent rolls of a three-sided die with weights that depend on r as shown in the following table (ignore the final column for the moment).

If r is in the interval...	...then Pr(1)=	...then Pr(2)=	...then Pr(3)=	If r' is in the interval...
$(1/2, 1]$	1/2	1/3	1/6	$[1, 2)$
$(1, 3/2]$	1/6	1/2	1/3	$(2/3, 1]$
$(3/2, 2]$	1/3	1/6	1/2	$(1/2, 2/3]$

For instance, if r is in the interval $(1/2, 1]$, then the mad genius will report the results of three rolls of a three-sided die such that the probability of 1 is 1/2, the probability of 2 is 1/3, and the probability of 3 is 1/6.

Suppose the mad genius reports the outcomes 1, 2, and 3. This outcome has the same probability $(1/2)(1/3)(1/6) = 1/36$ under each of the hypotheses $H_1 : 1/2 < r \leq 1$, $H_2 : 1 < r \leq 3/2$, and $H_3 : 3/2 < r \leq 2$. It is your total relevant evidence, so (PRAFE) says to accept each of those hypotheses over each of the others to degree zero (i.e., to be neutral among them). Thus, (3B) says to prefer $(H_2$ or $H_3) : 1 < r \leq 2$ to $H_1 : 1/2 < r \leq 1$ to a degree greater than zero.

Now consider the ratio r' of wine to water instead of the ratio r of water to wine. Before the die roll, you have no information about r' except that it is in the interval $[1/2, 2)$. The table above gives the probability distributions for the die roll outcomes under each possible value of r'. The outcome $\{1, 2, 3\}$ has the same probability $(1/2)(1/3)(1/6) = 1/36$ under each of the hypotheses $H_1' : 1/2 \leq r' < 2/3$, $H_2' : 2/3 \leq r' < 1$, and $H_3' : 1 \leq r' < 2$. Moreover, it has the same probability under each of the hypotheses $H_1^* : 1/2 \leq r' < 1$, $H_2^* : 1 \leq r' < 3/2$, and $H_3^* : 3/2 \leq r' < 2$. For if $H_1^* : 1/2 \leq r' < 1$ is true, then either $1/2 \leq r' < 2/3$ or $2/3 \leq r' < 1$. Either way, the probably of the outcome is 1/36. Thus, if H_1^* is true, then the probably the outcome is 1/36. If $H_2^* : 1 \leq r' < 3/2$ is true, then $1 \leq r' < 2$, so the probably the outcome is 1/36. Likewise for H_3^*. The die roll is your total relevant evidence, so (PRAFE) says to accept each of the hypotheses H_1^*, H_2^*, and H_3^* over each of the others to degree zero (i.e., to be neutral among them). Thus, (3B) says to prefer (H_2^* or H_3^*) to H_1^* to a degree greater than zero.

But now we have violated (3A). $(H_2$ or $H_3) : 1 < r \leq 2$ is equivalent to $H_1^* : 1/2 \leq r' < 1$, and $H_1 : 1/2 < r \leq 1$ is equivalent to (H_2^* or H_3^*) $: 1 \leq r' < 2$. Yet we have accepted $(H_2$ or $H_3)$ over H_1 and (H_2^* or H_3^*) over H_1^*. Therefore, (PRAFE) can force you to violate either (3A) or (3B).

Discussion

Two objections to this proof are worth discussing. First, statisticians distinguish between "simple" and "complex" statistical hypotheses. A simple statistical hypothesis says that the data- generating mechanism follows a particular probability distribution. A complex statistical hypothesis is a disjunction of simple statistical hypotheses. The Law of Likelihood applies in the first instance only to simple statistical hypotheses. It might seem that the hypotheses we consider here are complex statistical hypotheses. After all, they are disjunctions of more specific hypotheses about the value of r. One could claim that for this reason (PRAFE), properly understood, does not apply to those hypotheses.

That response to the example will not work. The hypotheses we have considered are disjunctions of hypotheses that posit particular values for r. But H_1, H_2, H_3, H_1', H_2', H_3', H_2^*, and H_3^* are not disjunctions of hypotheses that posit different probability distributions for the outcome of the die roll. They are instead disjunctions of hypotheses that all imply the same probability distribution for the outcome of the die roll. A likelihoodist who denied that we can say $\Pr(A|H) = a$ because $\Pr(A|H_i) = a$ for all H_i in some partition of H would be in deep trouble. It is this assumption that allows us to ignore irrelevant partitions of our hypotheses, which can always (or at least virtually always) be found. (For instance, we routinely ignore the fact that the hypothesis H that a given coin is fair can be partitioned into the hypothesis H_1 that the coin is fair and the moon is made of green cheese and the hypothesis H_2 that the coin is fair and the moon is not made of green cheese. In the same way, we can ignore the fact that the hypothesis $1/2 < r \leq 1$, for instance, can be partitioned into hypotheses of the form $r = 1/2 + \epsilon$ for $0 < \epsilon \leq 1/2$.)

Now, H_1^* is a disjunction of hypotheses not all which posit the same probability distribution for the outcome of the die roll. But we can arrive at a likelihood

for H_1^* on $\{1,2,3\}$ in several ways. One way, used in the proof, is to interpret likelihoods as probabilities entailed by hypotheses and to use disjunction elimination to derive that H_3^* entails $\Pr(\{1,2,3\}) = 1/36$. But Bayesians and some likelihoodists want to interpret likelihoods as conditional probabilities.

There are at least two ways to get a likelihood for H_1^* under this interpretation. One way is to invoke a version of the law of total probability: if $B = (B_1 \text{ or } B_2)$, then $\Pr(A|B) = \Pr(A|B_1)\Pr(B_1|B) + \Pr(A|B_2)\Pr(B_2|B)$. Thus,

$$\begin{aligned}
\Pr(\{1,2,3\}|H_1^*) &= \Pr(\{1,2,3\}|H_1')\Pr(H_1'|H_1^*) + \Pr(\{1,2,3\}|H_2')\Pr(H_2'|H_1^*) \\
&= 1/36\Pr(H_1'|H_1^*) + 1/36\Pr(H_2'|H_1^*) \\
&= 1/36[\Pr(H_1'|H_1^*) + \Pr(H_2'|H_1^*)] \\
&= 1/36\Pr(H_1' \text{ or } H_2'|H_1^*) \\
&= 1/36\Pr(H_1^*|H_1^*) \\
&= 1/36
\end{aligned}$$

Now, some likelihoodists would reject this argument. The result does not depend on the values of $\Pr(H_1'|H_1^*)$ and $\Pr(H_2'|H_1^*)$, but the argument does mention those values and thus assumes that they exist. Some likelihoodists would reject that assumption.

For a likelihoodist who interprets likelihoods as conditional probabilities and rejects the existence of probabilities that are not objectively well-defined, there is still another way to get a likelihood for H_1^*: invoke the Principal Principle. The Principal Principle (Lewis, 1981) says that one's credence for A given a proposition which entails that the chance of A is x and no inadmissible infor-

mation[16] should be x: $\mathrm{Cr}(A|H) = x$ where H entails $\mathrm{Ch}(A) = x$ and contains no inadmissible information. If we interpret $\Pr(\{1,2,3\}|H_1^*)$ as a credence, then it follows that $\Pr(\{1,2,3\}|H_1^*) = \Pr(\{1,2,3\}|H_1' \text{ or } H_2') = 1/36$, because $(H_1' \text{ or } H_2')$ entails $\mathrm{Ch}(\{1,2,3\}) = 1/36$ by a disjunction elimination. If we interpret $\Pr(\{1,2,3\}|H_1^*)$ as a chance rather than a credence, then we do not need the Principal Principle but only the transparently obvious chance-chance principle which says that $\mathrm{Ch}(A|H) = x$ where H entails $\mathrm{Ch}(A) = x$ and contains no inadmissible information.

Second, one could claim that the outcomes of the die rolls are not part of one's total relevant evidence with respect to the hypotheses under consideration. One's total relevant evidence with respect to those hypotheses is the empty set. After all, one's assessment of those hypotheses makes no difference to one's assessment of the probability of that outcome. This claim is somewhat reasonable, but it does not help. We could simply ask what the probability is of one's total relevant evidence with respect to the hypotheses under consideration in the empty set is under each of those hypotheses. For each hypothesis, that probability is simply the probability that the outcome of the die roll is 1, putting us back where we started.

Now, a natural response to this maneuver is to claim that (PRAFE) applies only to non-empty bodies of relevant evidence (or to non-neutral bodies of evidence if evidence can be both relevant and neutral). One could, for instance, adapt the approach to ignorance that Norton (2008) develops by assigning the same non-numerical degree of relative acceptance I to all pairs of contingent hypotheses in cases of neutral evidence. However, this response faces at least two difficulties. First, it creates unnatural discontinuities. On this proposal, we are not required to formulate any commitments about the hypotheses H_1,

[16] See (Lewis, 1981) for a discussion of admissibility. Roughly speaking, information is in admissible if it speaks to the *outcome* of a chance process rather than to its stochastic properties.

H_2, and H_3 in light of the evidence in the example as it stands—not even a commitment of neutrality. But suppose we modified the example slightly by adding some tiny quantity ϵ to the probability of 1 under each possible value of r and subtracting it from the probability of 3 under each possible value of r. Then we would be required to formulate commitments, namely minute preferences for H_2 over H_1 and over H_3. The idea that neutral evidence requires no commitments while arbitrarily slightly non-neutral evidence requires completely definite commitments is hard to accept.

Second, this response sets up a game of cat-and-mouse that seems unlikely to end well for the methodological likelihoodist. An example involving exactly neutral evidence is needed to illustrate a conflict among (PRAFE), (3A), and (3B) only because I made those principles very weak so that they would command nearly universal assent. Similar but slightly stronger principles would conflict in cases of non-neutral evidence. For instance, (3B) says to accept (H_1 or H_2) over H_3 to a degree greater than that to which you accept H_1 over H_3 under the relevant conditions. Presumably, in each case in which (3B) applies, there is some definite amount by which the former should exceed the latter. In the case at hand, for instance, there is some positive number t such that it is at least permissible to accept (H_2 or H_3) over H_1 to degree t. One could use this margin t to generate the same kind of argument in a similar case with sufficiently slightly non-neutral evidence.

B Proof that (PRAFE) can force one to violate (4A)

(4A) Do not prefer H_1 to $\sim H_1$ and $\sim H_2$ to H_2 if H_1 and H_2 are logically equivalent given your total evidence.

Proof

Let X_1 and X_2 record the outcomes of independent coin flips. If the first coin lands heads, then $X_1 = 1$. Otherwise $X_1 = 0$. Likewise, if the first coin lands heads, then $X_2 = 1$. Otherwise $X_2 = 0$. Let E be the evidence $X_1 = 1$, H_1 the hypothesis $X_1 = X_2 = 1$, and H_2 the hypothesis $X_2 = 1$. Suppose that E is the only information one has about X_1 and X_2.

Fix the function f such that (PRAFE) says to accept H_1 over H_2 to degree $f(\mathcal{L}) = f(\Pr(T|H_1)/\Pr(T|H_2))$. By the formulation of (PRAFE), there's some constant a such that $f(a) > 0$ (and thus $f(x) > 0$ for all $x > a$, since f is nondecreasing). I will show in a moment the following:

> (*) For any a, there is a joint distribution for X_1 and X_2 such that $\Pr(E|H_1)/\Pr(E| \sim H_1) > a$ and $\Pr(E| \sim H_2)/\Pr(E|H_2) > a$.

Thus, (PRAFE) forces one to prefer H_1 to $\sim H_1$ and $\sim H_2$ to H_2. But given E, H_1 and H_2 are equivalent. Therefore, (PRAFE) forces one to violate (4A).

I will now prove (*) by showing how to construct for any a a joint distribution over X_1 and X_2 such that such that $\Pr(E|H_1)/\Pr(E| \sim H_1) > a$ and $\Pr(E| \sim H_2)/\Pr(E|H_2) > a$. Let a be some value greater than $1/2$ of x such that $f(a) > 0$ for all $x > a$. Choose a $b > (2a-1)/(2a+1)$. Then assign probabilities to outcomes according to the following table.

$X_1 \backslash X_2$	0	1	
0	$\frac{1-b}{4}$	b	$\frac{1+3b}{4}$
1	$\frac{1-b}{4}$	$\frac{1-b}{2}$	$\frac{3-3b}{4}$
	$\frac{1-b}{2}$	$\frac{1+b}{2}$	1

Here is a derivation of the relevant likelihood ratios.

$$\begin{aligned}
\frac{\Pr(E|H_1)}{\Pr(E|\sim H_1)} &= \frac{\Pr(E\ \&\ H_1)}{Pr(H_1)}\frac{\Pr(\sim H_1)}{\Pr(E\ \&\ \sim H_1)} \\
&= \frac{\Pr(X_1 = X_2 = 1)}{\Pr(X_1 = X_2 = 1)}\frac{\Pr(\sim (X_1 = X_2 = 1))}{\Pr(X_1 = 1\ \&\ X_2 = 0)} \\
&= \frac{1 - \Pr(X_1 = X_2 = 1)}{\Pr(X_1 = 1\ \&\ X_2 = 0)} \\
&= \frac{1 - (1-b)/2}{(1-b)/4} \\
&= \frac{4}{1-b} - 2 \\
&= \frac{4 - 2 + 2b}{1-b} \\
&= 2\frac{1+b}{1-b}
\end{aligned}$$

This quantity is monotonically increasing in b. Thus, it follows that if $b > (2a-1)/(2a+1)$,[17] then

$$\begin{aligned}
\Pr(E|H_1)\Pr(E|\sim H_1) &> 2\frac{1 + (2a-1)/(2a+1)}{1 - (2a-1)/(2a+1)} \\
&= 2\frac{2a+1+2a-1}{2a+1-2a+1} \\
&= 2\frac{4a}{2} \\
&= 4a \\
&> a
\end{aligned}$$

And now the other likelihood ratio.

[17] $b > (2a-1)/(2a+1)$ is always permissible, because $a > 1/2$ implies $0 < (2a-1)/(2a+1) < 1$, and $0 < b < 1$ implies that the distribution in the table above is consistent with the axioms of probability.

$$\begin{aligned}
\frac{\Pr(E|\sim H_2)}{\Pr(E|H_2)} &= \frac{\Pr(E\ \&\ \sim H_2)}{\Pr(\sim H_2)}\frac{\Pr(H_2)}{\Pr(E\ \&\ H_2)} \\
&= \frac{\Pr(X_1 = 1\ \&\ X_2 = 0)}{\Pr(X_2 = 0)}\frac{\Pr(X_2 = 1)}{\Pr(X_1 = X_2 = 1)} \\
&= \frac{(1-b)/4}{(1-b)/2}\frac{(1+b)/2}{(1-b)/2} \\
&= 1/2\frac{1+b}{1-b}
\end{aligned}$$

This quantity is monotonically increasing in b. Thus, it follows that if $b > (2a-1)/(2a+1)$, then

$$\begin{aligned}
\frac{\Pr(E|\sim H_2)}{\Pr(E|H_2)} &> 1/2\frac{1+(2a-1)/(2a+1)}{1-(2a-1)/(2a+1)} \\
&= 1/2\frac{2a+1+2a-1}{2a+1-2a+1} \\
&= 1/2\frac{4a}{2} \\
&= a
\end{aligned}$$

Discussion

Note that restricting (PRAFE) so that it applies only to mutually exclusive hypotheses does not block this proof. (PRAFE) is applied only to the comparison between H_1 and $\sim H_1$ and the comparison between H_2 and $\sim H_2$. Those pairs of hypotheses are of course mutually exclusive. H_1 and H_2 are not mutually exclusive, but (PRAFE) is not applied to the comparison between H_1 and H_2 directly.

Title: Why (a Form of) Function Indeterminacy is Still a Problem for Biomedicine, and How Seeing Functional Items as Components of Mechanisms Can Solve it

Abstract: During the 1990s, many philosophers wrestled with the problem of function indeterminacy. Although interest in the problem has waned, I argue that solving the problem is of value for biomedical research and practice. This is because a solution to the problem is required in order to specify rigorously the conditions under which a given item is "dysfunctional." In the following I revisit a solution developed originally by Neander (1995), which uses functional analysis to solve the problem. I situate her solution in the framework of mechanistic explanation and suggest two improvements.

Keywords: Biological function; function indeterminacy; mechanistic explanation; philosophy of biology; philosophy of medicine

1. Biomedicine Needs a Solution to the Problem of Indeterminacy

The central organizing principle of biomedical intervention is that of *fixing dysfunctional items*. This is not to say that biomedical practitioners do not do other things besides fixing dysfunctions. Sometimes, instead of fixing dysfunctional items, practitioners simply remove those items from the body, or they supplement their activity, or they inhibit their activity so as to restore proper physiological functioning. Yet the idea of fixing dysfunctions is an organizing principle of biomedicine in the sense that it illuminates most of the other sorts of goals that biomedical practitioners have (with the exception of goals such as cosmetic surgery, or pain relief during labor). For example, when practitioners choose to remove, supplement, or inhibit dysfunctional items, rather than fix them, they typically do so because of various limitations on the ability to fix them. Moreover, the idea of repairing dysfunctions is also an organizing principle because it works as a heuristic for biomedical research. This is because researchers often do not feel that they entirely *understand* a pathological process until they know what they would need to do, in theory, in order to fix it.

Biomedicine is limited in its ability to fix dysfunctions because of various epistemic, technical, and sociopolitical obstacles. At least one of those obstacles, however, is conceptual, or, if you will, "metaphysical." The ability to carry out the ideal of fixing dysfunctions requires, in the first place, that we are able to clearly *articulate* the conditions under which an item is functional and the conditions under which it is dysfunctional. Yet this is precisely what, according to one version of the function indeterminacy problem, is precluded (e.g., Dretske 1986; Neander 1995). Consider a well-worn but lucid example: the heart beats. In doing so, it circulates the blood. In doing so, it brings nutrients to cells and removes waste. In doing so, it contributes to survival and, ultimately, to reproductive success. Yet which of these activities constitutes the *function* of the heart? Any one of them would be licensed by standard theories of biological function, and in particular, theories that tie function to selection history or current adaptiveness (see Garson 2015). (I will refer to both groups of theories as "evolutionary" theories of function because they tie function to evolutionary considerations, despite the fact that one set of theories focuses more on history and

another on present-day activity. I am not here concerned with Cummins-type (or "causal role") functions, in which the function of a trait simply consists in its contribution to some systemic capacity of interest to an investigator. This theory will come into play later. I will also justify this exclusion later.) The problem arises most clearly for evolutionary theories, though a version of the problem could arise for causal role theories as well. I will refer to this as the "hierarchical" version of the problem of functional indeterminacy for reasons to be explained in the next section.

Here is where the problem comes in for biomedicine. Suppose that one succumbs to the temptation of pluralism, and asserts that there is no *principled* and *context-independent* way of selecting one of those descriptions of the heart's function (e.g., in terms of beating, circulating blood, bringing nutrients to cells, etc.) as the uniquely correct description of its function. Any of those activities, one might hold, may constitute the heart's "function," depending on factors such as disciplinary interest, convention, or personal predilection. The pluralist solution runs into trouble when we realize that these different function ascriptions can conflict with one another. Specifically, it is possible for the heart to carry out one of these activities and not the other. For example, suppose that the heart beats, but, due to a massive brain hemorrhage, the heart cannot circulate enough blood to keep the individual alive. Should we say that the heart is failing to perform its function of circulating blood (or not at the appropriate rate)? Or should we say, instead, that the heart is functioning successfully because it is performing its function of beating, despite the fact that, due to the hemorrhage, the activity is not associated with its normal contribution to survival?

Intuitively – if you and I share the same intuitions – we should say that the heart is functioning. It just cannot make its normal contribution to fitness because some *other* item is dysfunctional, namely, the ruptured artery. After all, the heart is only "doing its job," but the artery isn't doing *its* job. Deciding whether or not the heart is functioning in this situation is like trying to locate blame in a large corporation. But in saying this, we are privileging one activity over another as having a greater claim to constituting the function of the heart: the claim that the function of the heart is to *beat* is "privileged" in a way that the claim that the function of the heart is to *circulate blood* is not. Moreover, it seems to be privileged in some principled, context-independent way. This is precisely what the pluralist solution forbids. So, how can we justify this assertion that the one function ascription ("the heart beats") is more correct than another ("the heart circulates blood")?

Keep in mind that our solution – that in the case of the hemorrhage, the heart is "functional" and not "dysfunctional" – is not *only* intuitively correct. To maintain otherwise would be counterproductive or contrary to the needs to biomedicine. This is because, when we say that an item is *dysfunctional*, we indicate that the item in question is a prime target for direct biomedical intervention, such as repair or replacement. But if the heart cannot circulate blood effectively because of a ruptured artery in the brain, we presumably want to fix the artery, not the heart!

As a consequence, any attempt to articulate clearly the conditions under which an item is dysfunctional entails, as a necessary condition, a solution to the problem of function indeterminacy. In other words, if one *can* articulate clearly the conditions under which an item is dysfunctional, then one can use those conditions to state, in a principled and context-independent way, which of the multiple function ascriptions legitimized by evolutionary considerations is uniquely correct (or at least one can reduce the plurality of reasonable function ascriptions down to a small number of equally correct ascriptions). Conversely, if one possesses a solution to the problem of function indeterminacy, then (presumably) one could apply that to resolve such conflicts in the biomedical context. How shall we proceed?

In the following I will do three things. First, I will revisit a solution proposed by Neander (1995). In short, in her view, in order to identify the (determinate) function of any given item, we can utilize the framework of functional analysis. The (determinate) function of an item is identified with its "most specific function," which turns out to be its causal role within a certain mechanism. Next, I will describe that solution using the framework of multi-level mechanistic explanation as it has been developed over the last two decades (Section 2). Finally, I suggest two improvements to that solution (Section 3), one minor and one more substantial. The first is to replace talk of the "most specific function" with the "differentiated function" of the item. The second is to draw attention to *two* dimensions of indeterminacy, a "horizontal" dimension and a "vertical" dimension, and to suggest how mechanistic modeling can resolve both types.

2. Seeing Functional Items as Components in a Nested Hierarchy of Mechanisms

The central idea of the solution to the problem of indeterminacy that I will present here advances Neander's (1995) solution by anchoring it more firmly within the literature of the new mechanism tradition (e.g., Bechtel and Richardson 2010; Glennan 1996, 2005; MDC 2000; Craver 2001; Darden 2006; Craver and Darden 2013). It then develops it in two ways. In short, we can solve the indeterminacy problem by construing the functional item as a component within a mechanism, or, more precisely, within a nested hierarchy of mechanisms.

More specifically, the hierarchical version of the problem of indeterminacy stems from the fact that for any given item and any given function, there is a hierarchy of activities that explains why the performance of the function is (or was) associated with some biological advantage for the organism. Quite fortunately – and this is the key to solving it – that hierarchy of activities is mirrored by a corresponding hierarchy of mechanisms. Moreover, these mechanisms are nested, one within the other. For example, suppose A and A′ are two activities "adjacent" to one another on the functional hierarchy, where A′ is "higher" than A (see Figure 1). Suppose A is the activity of beating, and A′ the activity of circulating blood. On the corresponding mechanistic hierarchy, there is a mechanism, M, for A, and another mechanism, M′, for A′. In this case (and simplifying tremendously), M is comprised of the heart, and M′ is comprised of the circulatory

system, which includes, in addition to the heart, the brain and blood vessels. M and M′ are nested in the sense that M can be construed as a component part of M′.

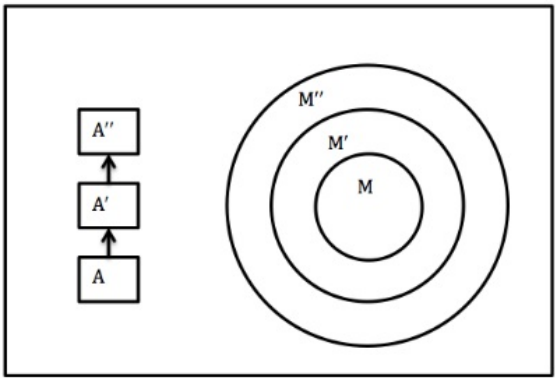


Figure 1. Functional and mechanistic hierarchies

With this framework in mind, we can solve the function indeterminacy problem by identifying the function of any particular item as *its differentiated contribution to the activity of the mechanism of which it is a component* – that is, the mechanism in which it is most "immediately" contained. For example, in this framework, it would be correct to say that the function of the heart is *to beat*, since that is its specific contribution to the activity A′ (blood circulation) of the mechanism M′ (circulatory system) in which it is "immediately" contained. To put the point differently, and somewhat more methodically: in order to identify the function of an item, we first construct our functional hierarchy. This hierarchy will be determined, in part, by our theory of function and in part by the empirical facts. We then construct a corresponding mechanistic hierarchy. Then, beginning at the uppermost level of the mechanistic hierarchy, and we work our way "downwards," level by level, until we reach the level at which the item in question emerges as an unanalyzed component (Neander 1995, 129). This is the "bottom-out" level as far as our analysis is concerned. Then the *indeterminate* function of the item can be rendered *determinate* by identifying it with the contribution that the item makes to the activity of the mechanism of which it is a component.

The same solution can, if necessary, be cast in a more historical vein, for example, in terms of the selected effects theory of function. The solution simply requires that instead of analyzing the trait's *current* contribution to fitness, we attempt to reconstruct, historically, how it may have contributed to fitness in the past. As noted above, I do not wish to take a stance, in this context, on *which* evolutionary approach to function is preferable, the approach that focuses on history or the approach that focuses on current-day performance (and, within each set of theories, which version of the theory is superior). The point is that the mechanistic framework could be adapted easily to suit the needs of either.

It would be natural at this point for one to wonder how this solution differs from the Cummins' type causal role theory of function (Cummins 1975), or its more recent variants, such as the mechanistic causal role theory of function (e.g., Craver 2001, 2013). In Cummins' theory, the function of a trait consists in its contribution to some systemic capacity picked out by an investigator. In the mechanistic causal role version of this theory, the function of a trait consists in its contribution, in tandem with the other parts of the system, to a systemic capacity of interest, and the whole system is described using the framework of mechanistic explanation. Isn't what I've presented *just* the solution that the Cummins-type theory, on its own, would entail?

The main difference is that in the view presented here, the function of the item is first identified by utilizing an evolutionary framework, and specifically, by considering the item in light of its selection history or its current contribution to fitness. That evolutionary framework provides a rationale for selecting a certain hierarchy of activities (pumping -> circulating blood -> bringing nutrients to cells -> helping creature survive and reproduce) as constituting the heart's (indeterminate) function. (Of course, which specific sequence of activities we identify depends partly on which specific theory of function we select within that family of theories, a topic on which, again, I wish to remain neutral at present.) A problem with this solution, of course, is that it creates the hierarchical version of the problem of function indeterminacy. We *then* apply a mechanistic framework to resolve that problem in a principled way. The hierarchy of functions that has been identified by evolutionary considerations is converted into a framework for identifying the relevant hierarchy of mechanisms, and we then look for the function of the item *qua* mechanistic causal role. Mechanistic considerations merely help us to make the transition from an indeterminately-specified function to a determinately-specified one. They do not supplant evolutionary considerations.

Even if this solution differs from the causal role theory of functions, one might wonder why the solution offered here is preferable. Why not just drop the evolutionary framework and go straightaway to a Cummins-type framework? The reason is that utilizing the evolutionary framework allows us to avoid certain recalcitrant problems associated with the Cummins-type theory, in particular, the problems of overbreadth and normativity. The first is the classic problem of overbreadth (e.g., Millikan 1989, 294; Kitcher 1993, 390). In Cummins' view, as in the mechanistic causal role view, the choice of a "top level" function for a given system is largely a matter of caprice. (I mean this in the sense that there is no objective, mind-independent fact of the matter regarding what the "top-level" function of any given system is, and not in the sense that it is somehow unmotivated or unjustifiable in any given case.) A consequence of this is that Cummins' framework licenses wildly counterintuitive function ascriptions (for example, Cummins' (1975, 752) own example that the function of the appendix could be to produce appendicitis. Incidentally, he raises this as a problem for Nagel's theory, but does not suggest how his own theory would resolve it). Alternatively, theories of function that tie function to evolutionary considerations have the implication that there *is* an objective fact of the matter regarding what the function of a given item (albeit indeterminate) actually

is, and it yields function ascriptions that are (typically) in line with biologically-informed intuitions. At least some philosophers find that to be a welcome implication.

Additionally, Cummins-type theories have a notoriously difficult time explaining the *normativity* of functions, by which I mean the fact that it is possible for something to possess a function without being able to perform that function (e.g., Neander 1991, 181-2; though see Hardcastle 2002; Cummins et al. 2010). That is because, on Cummins' view, the function of a trait is a *disposition* (Cummins 1975, 758). If a trait loses the disposition to perform a certain activity, then, at least on the classic view, it loses the function itself. In that case, how can a trait dysfunction? These two problems (overbreadth and normativity) provide some motivation for retaining an evolutionary perspective for thinking about function, but supplementing, rather than supplanting, that perspective with considerations drawn from mechanistic explanation.

3. Hierarchical and Sequential Aspects of Function Indeterminacy

Up until this point, I have largely described Neander's solution, or re-described it slightly. Yet I propose to extend that solution in two ways. The first is fairly minor and the second more substantial. First, for Neander, the correct way of describing the function of an item (in cases of conflict, e.g., in the biomedical context) is in terms of what she calls the item's "most specific function." But I would prefer to speak of the item's "differentiated function" (which is to be distinguished from the developmental process of "differentiation," as in, e.g., cell differentiation). The phrase, "most specific function," contains an ambiguity that can be clarified using the framework of mechanistic explanation.

As she recognizes, the "most specific function" of an item can be described in at least two ways (Neander 1995, 118-119). From one perspective, we can describe the activity that the item produces without indicating *how* the activity contributes to the mechanism of which it is a part (for example, "the function of the heart is to beat"). From another perspective, we can merely indicate *that* the activity contributes to the activity of the mechanism of which the item is a part, without specifying the "intrinsic" nature of the activity (for example, "the function of the heart is to help circulate blood"). Craver (2001, 65) makes the same distinction, and describes these in terms of the "isolated activity" and "contextual role" of an item. It seems to me that, from the biomedical perspective, the former ascription ("the heart beats") is the more informative of the two, because the latter is overly generic: it does not differentiate the function of the heart from that of the other components of the circulatory system. Moreover, since we are envisioning the heart as a component within a mechanism for circulating blood, the fact that the heart contributes to blood circulation will be implicit in the models that we use to represent the mechanism. Instead of describing the item's "most specific function," then, I will describe the item's "differentiated function." Talking of the "differentiated function" of the heart brings to the front and center of attention that the heart is part of a *system* in which each part has a distinct, and different, causal role. It draws attention to what the heart does that *differs*

from what the other components do. The "differentiated function" of the heart, for example, is to beat, not to help circulate blood.

Secondly, Neander describes the "most specific function" of the item as the activity that it can perform "more or less on its own", rather than "in collaboration with other components" (118). Applied to the heart, the idea would be that the most specific function of the heart is to *beat*, rather than to *circulate blood* or to *deliver nutrients to cells*. However, in what sense is it true to say that the heart beats "more or less on its own?" Suppose that the function of the heart is to *beat*, but the heart stops beating, or stops beating at the appropriate rate, because, due to a lesion in the medulla, it is not receiving the proper impulses from the brain. Is the heart dysfunctional? Intuitively, it is not dysfunctional; rather, it has been placed in an abnormal circumstance in which it cannot perform its function. It has, as it were, merely been deprived of the right inputs, similar to an unplugged electrical toy. But it seems both counterintuitive, and counterproductive, to say that the heart is dysfunctional if it stops beating just because it is not receiving the right inputs. (Of course, there is one sense in which the heart beats "more or less on its own," namely, that in which it may continue to beat very briefly after removal from the body, as after pithing a frog. But that phenomenon is pretty short-lived!)

So, there is no obvious sense in which *beating* is something the heart can do more or less on its own. It requires the right sorts of inputs from other sources. This suggests that, when we are attempting to articulate clearly the conditions under which an item is dysfunctional, it is not enough simply to point to its causal role within a *hierarchy* of mechanisms. Rather, we must also provide at least a "mechanism sketch" of the way in which the item interacts with others, at the same level, to produce the activity of the mechanism of which it is a part. We must adopt, not only a hierarchical (or vertical) perspective, but a sequential (or horizontal) perspective on the mechanism as well. In order to state clearly the conditions under which an item is dysfunctional, that item's performing a function (e.g., the heart's beating) must be seen as *one stage in a productive sequence of activities* that are collectively responsible for yielding the activity of the mechanism as a whole (blood circulation). In light of such a mechanism sketch, we have the tools to specify that the item is question is dysfunctional *not only* when it cannot perform its differentiated function, but when it cannot perform its differentiated function *even when the other parts of the system have performed their own characteristic activities in their appropriate sequence* (see Garson and Piccinini 2014). Thus, a full specification of the conditions under which an item is dysfunctional can be made so long as we have some characterization of both the hierarchical and the sequential aspects of the mechanism in which the item is embedded.

To give a simple example: suppose we want to provide a mechanistic explanation for how the gut digests food. We would analyze it into several parts, such as the mouth, tongue, esophagus, stomach, small and large intestine, and anus, each with its differentiated function. Using the solution to the indeterminacy problem developed here, we could say that the function of the stomach is to break down food and transfer it to the duodenum. (Note that the stomach has several functions, for example, to protect the inner

organs from the highly corrosive acids it contains. Nothing in the solution to functional indeterminacy precludes the possibility that one trait or organ possesses several distinct functions, each associated with its own indeterministic "hierarchy.") Suppose, however, at a given moment, the stomach is not digesting any food. That does not mean that it is dysfunctional. After all, it is possible that the animal is fasting and there is no food to digest. Functions are "situation specific" (Kingma 2010). It is only dysfunctional if it is not breaking down food *and* all of the preceding stages (e.g., functions) in the sequence of digestion have taken place (culminating with food being delivered by the esophagus to the stomach).

A concern one might have with the introduction of this "horizontal" approach to defining dysfunction is that it appears, on the surface, to involve circularity. That is, we are trying to explain how it is that a trait can dysfunction (e.g., the stomach) and in so doing, we are appealing to the functions of the other parts of the system (e.g., the fact that the esophagus has discharged its function of bringing food to the stomach). But there is no circularity here. We first use the "vertical" approach to identify the *determinate* function of any given trait. That approach does not involve any apparent circularity because in order for me to identify the (determinate) function of the stomach (namely, to pass food along to the duodenum), I don't have to have already identified the (determinate) functions of the other parts of the system. It is enough that I have identified their indeterminate functions.

Once we have used that vertical method to identify the determinate functions of several components of a system, we can *then* deploy the "horizontal" perspective to identify precisely *the conditions under which any given component is dysfunctional.* At this stage (that is, in trying to understand when a part of the system is dysfunctional) we are free to make use of our knowledge of the determinate functions of the other parts of the system. What we are *not* allowed to do is to identify the determinate function of a component of a system by appealing to the determinate functions of the other parts of the system. What we are also not allowed to do is to identify the conditions under which an item is dysfunctional by appealing to the conditions under which some other item is dysfunctional. The approach I have outlined here avoids both sorts of circularity. In this way, the tools of mechanistic explanation help solve both of these aspects of the indeterminacy problem and allow us to state clearly when a given item is dysfunctional. More generally, this analysis suggests the importance of philosophical work on integrating considerations drawn from the traditional body of philosophical literature on biological function, and those drawn from the philosophical literature on mechanism.

References

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Craver, C. 2001 "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68:53–74.

Craver, C. 2013. **"**Functions and Mechanisms: A Perspectivalist View.**"** In *Functions: Selection and Mechanisms*, ed. Philippe Huneman, 133–58. Dordrecht: Springer.

Craver, C., and Darden, L. 2013. *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago: University of Chicago Press.

Cummins, R. 1975. "Functional Analysis." *Journal of Philosophy* 72: 741–765.

Cummins, R., and Roth, M. 2010. "Traits Have Not Evolved to Function the Way They Do Because of a Past Advantage." In *Contemporary Debates in Philosophy of Biology*, eds. F. Ayala and R. Arp, 72–86. Oxford: Blackwell.

Darden, Lindley. 2006. *Reasoning in Biological Discoveries*. Cambridge: Cambridge University Press.

Dretske, Fred. 1986. **"**Misrepresentation.**"** In *Belief: Form, Content, and Function*, ed. R. Bogdan, 17–36. Oxford: Clarendon.

Garson, Justin. 2015. *The Biological Mind: A Philosophical Introduction*. London: Routledge.

Garson, J., and Piccinini, G. 2014. "Functions Must be Performed at Appropriate Rates in Appropriate Situations." *British Journal for the Philosophy of Science* 65:1-20.

Glennan, Stuart. 1996. **"**Mechanisms and the Nature of Causation.**"** *Erkenntnis* 44:49–71.

Glennan, Stuart. 2005. **"**Modeling Mechanisms.**"** *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:443–64.

Hardcastle, V. G. 2002. **"**On the Normativity of Functions.**"** In *Functions: New Essays in the Philosophy of Psychology and Biology,* ed. Andre Ariew, Robert Cummins, and Mark Perlman, 144–56. Oxford: Oxford University Press.

Kingma, E. 2010. "Paracetamol, Poison, and Polio: Why Boorse's Account of Function Fails to Distinguish Health and Disease." *British Journal for the Philosophy of Science* 61:241–264.

Kitcher, P. 1993. "Function and design." *Midwest Studies in Philosophy* 18:379–397.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. **"**Thinking about Mechanisms.**"** *Philosophy of Science* 67:1–25.

Millikan, R.G. 1989. "In Defense of Proper Functions." *Philosophy of Science* 56:288–302.

Neander, K. 1991. "Functions as Selected Effects: The Conceptual Analyst's Defense." *Philosophy of Science* 58:168–184.

Neander, K. 1995. "Misrepresenting and Malfunctioning." *Philosophical Studies* 79:109-141.

PSA 2014: 24th Biennial Meeting of the Philosophy of Science Association (Chicago, IL)

How to Move an Electromagnetic Field?

Márton Gömöri and László E. Szabó
Department of Logic, Institute of Philosophy
Eötvös University, Budapest

Abstract

The special relativity principle presupposes that the states of the physical system concerned can be meaningfully characterized, at least locally, as such in which the system is at rest or in motion with some velocity relative to an arbitrary frame of reference. In the first part of the paper we show that electrodynamic systems, in general, do not satisfy this condition. In the second part of the paper we argue that exactly the same condition serves as a necessary condition for the persistence of an extended physical object. As a consequence, we argue, electromagnetic field strengths cannot be the individuating properties of electromagnetic field—contrary to the standard realistic interpretation of CED. In other words, CED is ontologically incomplete.

1 Introduction

The problem we address in this paper is on the border-line between physics and metaphysics. We begin with the observation that the special relativity principle (RP) is about the comparison of the behaviors of physical systems in different states of inertial *motion* relative to an arbitrary inertial frame of reference. Therefore, it is a minimal requirement for the RP to be a meaningful statement that the states of the system in question must be meaningfully characterized as such in which the system as a whole is at rest or in motion with some velocity relative to an arbitrary frame of reference. Thus, to apply the RP to classical electrodynamics (CED), it has to be meaningfully formulated when an electrodynamic system—charged particles plus electromagnetic field—is at rest or in motion relative to an inertial frame of reference. In the first part of the paper we formulate a minimal condition a solution of the Maxwell–Lorentz equations must satisfy in order to describe such an electrodynamic configuration. Then we prove that the solutions of the Maxwell–Lorentz equations, in general, do not satisfy these conditions.

In the second part of the paper, we discuss the conceptual relationship between the problem of motion and the problem of persistence. We argue that persistence presupposes—zero or non-zero—velocity. One can formulate a necessary condition for the persistence of an object, in terms of its individuating properties. This condition implies that the object must be in motion with some instantaneous velocity; or, in case of an extended object, its local parts must be in motion with some local and instantaneous velocities. At this point the problem of persistence connects to the problem discussed in the first part of the paper. As it is proved in Section 3, electromagnetic field does not satisfy this condition. Therefore, we conclude, electromagnetic field cannot be regarded as a real physical entity persisting in space and time; or, the field strengths cannot be regarded as fundamental quantities individuating electromagnetic field, that is, electrodynamics cannot be regarded as an ontologically complete description of electromagnetic phenomena.

2 The RP Is about the Behaviors of Physical Systems in Different States of Motion

The RP is one of the fundamental principles which must be satisfied by all laws of physics describing any physical phenomena. Without entering into the more technical formulation of the principle (see e.g. Gömöri and Szabó 2013), we would like to focus on one particular aspect, which is already clearly there in Galileo's first formulation:

> Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a wide vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to your friend, you need throw it no more strongly in one direction than another, the distances being equal; jumping with your feet together, you pass equal spaces in every direction. When you have observed all these things carefully (though doubtless when the ship is standing still everything must happen in this way), have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still. In jumping, you will pass on the floor the same spaces as before, nor will you make larger jumps toward the stern than toward the prow even though the ship is moving quite rapidly, despite the fact that during the time that you are in the air the floor under you will be going in a direction opposite to your jump. In throwing something to your companion, you will need no more force to get it to him whether he is in the direction of the bow or the stern, with yourself situated opposite. The droplets will fall as before into the vessel beneath without dropping toward the stern,

> although while the drops are in the air the ship runs many spans. The fish in their water will swim toward the front of their bowl with no more effort than toward the back, and will go with equal ease to bait placed anywhere around the edges of the bowl. Finally the butterflies and flies will continue their flights indifferently toward every side, nor will it ever happen that they are concentrated toward the stern, as if tired out from keeping up with the course of the ship, from which they will have been separated during long intervals by keeping themselves in the air. And if smoke is made by burning some incense, it will be seen going up in the form of a little cloud, remaining still and moving no more toward one side than the other. The cause of all these correspondences of effects is the fact that *the ship's motion is common to all the things contained in it* [italics added], and to the air also. That is why I said you should be below decks; for if this took place above in the open air, which would not follow the course of the ship, more or less noticeable differences would be seen in some of the effects noted. (Galilei 1953, 187)

What is important for our present concern is that the principle is about the comparison of the behaviors of physical systems—flies, butterflies, fishes, droplets, smoke—*in different states of inertial motion* relative to an arbitrary inertial frame of reference. In Brown's words:

> The principle compares the outcome of relevant processes inside the cabin under different states of inertial motion of the cabin relative to the shore. It is simply assumed by Galileo that the same initial conditions in the cabin can always be reproduced. What gives the relativity principle empirical content is the fact that the differing states of motion of the cabin are clearly distinguishable relative to the earth's rest frame. (Brown 2005, 34)

The RP describes the relationship between two situations: one is in which the system, as a whole, is at rest relative to one inertial frame, say K, the other is in which the system shows the similar behavior, but being in a collective motion relative to K, co-moving with some K'. In other words, the RP assigns to each solution F of the physical equations, stipulated to describe the situation in which the system is co-moving as a whole with inertial frame K, another solution $M_{\mathbf{V}}(F)$, describing the similar behavior of the same system when it is, as a whole, co-moving with inertial frame K', that is, when it is in a collective motion with velocity $\mathbf{V}$ relative to K, where $\mathbf{V}$ is the velocity of K' relative to K. And it asserts that the solution $M_{\mathbf{V}}(F)$, expressed in the primed variables of K', has exactly the same form as F in the original variables of K.

Consequently, the following is a minimal requirement for the RP to be a meaningful statement:

Minimal Requirement for the RP (MR) *The states of the system in question—described by the solutions F—must be meaningfully characterized as such in which the system as a whole is at rest or in motion with some velocity relative to an arbitrary frame of reference.*

Let us show a well-known electrodynamic example in which a particles + electromagnetic field system satisfies this condition. Consider one single charged

particle moving with constant velocity $\mathbf{V} = (V, 0, 0)$ relative to K and the coupled stationary electromagnetic field (Jackson 1999, 661):

$$M_{\mathbf{V}}(F) \begin{cases} E_x(x,y,z,t) = \dfrac{qX_0}{\left(X_0^2 + (y-y_0)^2 + (z-z_0)^2\right)^{3/2}} \\ E_y(x,y,z,t) = \dfrac{\gamma q\,(y-y_0)}{\left(X_0^2 + (y-y_0)^2 + (z-z_0)^2\right)^{3/2}} \\ E_z(x,y,z,t) = \dfrac{\gamma q\,(z-z_0)}{\left(X_0^2 + (y-y_0)^2 + (z-z_0)^2\right)^{3/2}} \\ B_x(x,y,z,t) = 0 \\ B_y(x,y,z,t) = -c^{-2}VE_z \\ B_z(x,y,z,t) = c^{-2}VE_y \\ \varrho(x,y,z,t) = q\delta\left(x - (x_0 + Vt)\right)\delta\left(y - y_0\right)\delta\left(z - z_0\right) \end{cases} \tag{1}$$

where (x_0, y_0, z_0) is the initial position of the particle at $t = 0$, $X_0 = \gamma\left(x - (x_0 + Vt)\right)$ and $\gamma = \left(1 - \frac{V^2}{c^2}\right)^{-\frac{1}{2}}$. In this case, it is no problem to characterize the particle + electromagnetic field system as such which is, as a whole, in motion with velocity $\mathbf{V}$ relative to K; as the electromagnetic field is in collective motion with the point charge of velocity $\mathbf{V}$ (Fig. 1) in the following sense:[1]

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r} - \mathbf{V}\delta t, t - \delta t) \tag{2}$$

$$\mathbf{B}(\mathbf{r}, t) = \mathbf{B}(\mathbf{r} - \mathbf{V}\delta t, t - \delta t) \tag{3}$$

that is,

$$-\partial_t \mathbf{E}(\mathbf{r}, t) = \mathsf{D}\mathbf{E}(\mathbf{r}, t)\mathbf{V} \tag{4}$$

$$-\partial_t \mathbf{B}(\mathbf{r}, t) = \mathsf{D}\mathbf{B}(\mathbf{r}, t)\mathbf{V} \tag{5}$$

where $\mathsf{D}\mathbf{E}(\mathbf{r}, t)$ and $\mathsf{D}\mathbf{B}(\mathbf{r}, t)$ denote the spatial derivative operators (Jacobians

[1] It must be pointed out that velocity $\mathbf{V}$ conceptually differs from the speed of light c. Basically, c is a constant of nature in the Maxwell–Lorentz equations, which can emerge in the solutions of the equations; and, in some cases, it can be interpreted as the velocity of propagation of changes in the electromagnetic field. For example, in our case, the stationary field of a uniformly moving point charge, in collective motion with velocity $\mathbf{V}$, can be constructed from the superposition of retarded potentials, in which the retardation is calculated with velocity c; nevertheless, the two velocities are different concepts. To illustrate the difference, consider the fields of a charge at rest (9), and in motion (1). The speed of light c plays the same role in both cases. Both fields can be constructed from the superposition of retarded potentials in which the retardation is calculated with velocity c. Also, in both cases, a small local perturbation in the field configuration would propagate with velocity c. But still, there is a consensus to say that the system described by (9) is at rest while the one described by (1) is moving with velocity $\mathbf{V}$ (together with K', relative to K.) A good analogy would be a Lorentz contracted moving rod: $\mathbf{V}$ is the velocity of the rod, which differs from the speed of sound in the rod.

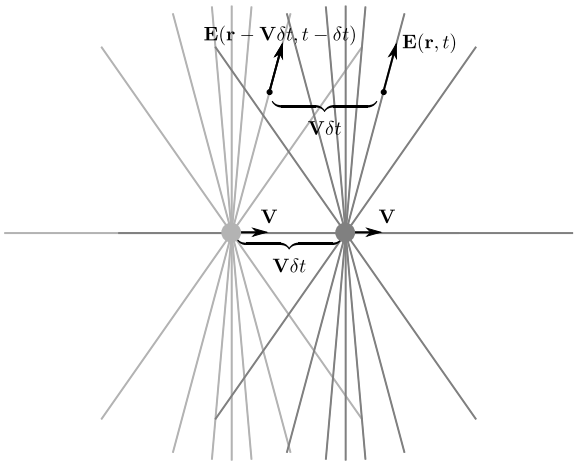


Figure 1: The stationary field of a uniformly moving point charge is in collective motion together with the point charge

for variables x, y and z); that is, in components:

$$\begin{aligned}
-\partial_t E_x(\mathbf{r},t) &= V_x\partial_x E_x(\mathbf{r},t) + V_y\partial_y E_x(\mathbf{r},t) + V_z\partial_z E_x(\mathbf{r},t) && (6)\\
-\partial_t E_y(\mathbf{r},t) &= V_x\partial_x E_y(\mathbf{r},t) + V_y\partial_y E_y(\mathbf{r},t) + V_z\partial_z E_y(\mathbf{r},t) && (7)\\
&\vdots\\
-\partial_t B_z(\mathbf{r},t) &= V_x\partial_x B_z(\mathbf{r},t) + V_y\partial_y B_z(\mathbf{r},t) + V_z\partial_z B_z(\mathbf{r},t) && (8)
\end{aligned}$$

The uniformly moving point charge + electromagnetic field system not only satisfies condition MR, but it satisfies the RP: Formula (1) with $\mathbf{V} = 0$ describes the static field of the particle when they are at rest in K :

$$F\begin{cases}
E_x(x,y,z,t) = \dfrac{q\,(x-x_0)}{\left((x-x_0)^2+(y-y_0)^2+(z-z_0)^2\right)^{3/2}}\\
E_y(x,y,z,t) = \dfrac{q\,(y-y_0)}{\left((x-x_0)^2+(y-y_0)^2+(z-z_0)^2\right)^{3/2}}\\
E_z(x,y,z,t) = \dfrac{q\,(z-z_0)}{\left((x-x_0)^2+(y-y_0)^2+(z-z_0)^2\right)^{3/2}}\\
B_x(x,y,z,t) = 0\\
B_y(x,y,z,t) = 0\\
B_z(x,y,z,t) = 0\\
\varrho(x,y,z,t) = q\delta\,(x-x_0)\,\delta\,(y-y_0)\,\delta\,(z-z_0)
\end{cases} \qquad (9)$$

By means of the Lorentz transformation rules one can express (1) in terms of

the 'primed' variables of the co-moving reference frame K':

$$\begin{aligned}
E'_x(x',y',z',t') &= \frac{q'\,(x'-x'_0)}{\left((x'-x'_0)^2+(y'-y'_0)^2+(z'-z'_0)^2\right)^{3/2}} \\
E'_y(x',y',z',t') &= \frac{q'\,(y'-y'_0)}{\left((x'-x'_0)^2+(y'-y'_0)^2+(z'-z'_0)^2\right)^{3/2}} \\
E'_z(x',y',z',t') &= \frac{q'\,(z'-z'_0)}{\left((x'-x'_0)^2+(y'-y'_0)^2+(z'-z'_0)^2\right)^{3/2}} \\
B'_x(x',y',z',t') &= 0 \\
B'_y(x',y',z',t') &= 0 \\
B'_z(x',y',z',t') &= 0 \\
\varrho(x',y',z',t') &= q\delta\,(x'-x'_0)\,\delta\,(y'-y'_0)\,\delta\,(z'-z'_0)
\end{aligned} \tag{10}$$

and we find that the result is indeed of the same form as (9).

So, in this well-known particular textbook example the RP is meaningful and satisfied. This picture is in complete accordance with the standard realistic interpretation of electromagnetic field:

> In the standard interpretation of the formalism, the field strengths **B** and **E** are interpreted realistically: The interaction between charged particles are mediated by the electromagnetic field, which is ontologically on a par with charged particles and the state of which is given by the values of the field strengths. (Frisch 2005, 28)

In this example, the charged particle and the coupled electromagnetic field constitute a physical system which—just like Galileo's flies, butterflies, fishes, droplets, and smoke—can be subject to the RP. The states F and $M_{\mathbf{V}}(F)$ can be meaningfully characterized as such in which both parts of the physical system, the particle and the electromagnetic field, are at rest or in motion with some velocity relative to an arbitrary frame of reference. We will show, however, that this is not the case in general.

3 How to Understand the RP for a General Electrodynamic System?

What meaning can be attached to the words "a coupled particles + electromagnetic field system is *in collective motion* with velocity $\mathbf{V}$" ($\mathbf{V}=0$ included) relative to a reference frame K, in general? One might think, we can read off the answer to this question from the above example. However, focusing on the electromagnetic field, the partial differential equations (4)–(5) imply that

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0(\mathbf{r}-\mathbf{V}t) \tag{11}$$

$$\mathbf{B}(\mathbf{r},t) = \mathbf{B}_0(\mathbf{r}-\mathbf{V}t) \tag{12}$$

with some time-independent $\mathbf{E}_0(\mathbf{r})$ and $\mathbf{B}_0(\mathbf{r})$. In other words, the field must be a stationary one, that is, a translation of a static field with velocity $\mathbf{V}$. But,

(11)–(12) is certainly not the case for a general solution of the equations of CED; the field is not necessarily translating with a collective velocity. The behavior of the field can be much more complex. Whatever this complex behavior is, it is quite intuitive to assume that the following general principle must hold:

Mereological Principle of Motion (MPM) *If an extended object as a whole is at rest or is in motion with some velocity relative to an arbitrary reference frame K, then all local parts of it are in motion with some local instantaneous velocity $\mathbf{v}(\mathbf{r},t)$ relative to K.*

Combining MPM with MR, we obtain the following:

Local Minimal Requirement for the RP (LMR) *The states of the extended physical system in question must be meaningfully characterized as such in which all local parts of the system are at rest or in motion with some local instantaneous velocity relative to an arbitrary frame of reference.*

Consequently, in case of electrodynamics, a straightforward minimal requirement for the RP to be a meaningful statement is that (2)–(3) must be satisfied at least *locally* with some local and instantaneous velocity $\mathbf{v}(\mathbf{r},t)$: it is quite natural to say that the electromagnetic field at point $\mathbf{r}$ and time t is moving with *local* and *instantaneous* velocity $\mathbf{v}(\mathbf{r},t)$ if and only if

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}\left(\mathbf{r}-\mathbf{v}(\mathbf{r},t)\delta t, t-\delta t\right) \tag{13}$$
$$\mathbf{B}(\mathbf{r},t) = \mathbf{B}\left(\mathbf{r}-\mathbf{v}(\mathbf{r},t)\delta t, t-\delta t\right) \tag{14}$$

are satisfied *locally,* in an *infinitesimally* small space and time region at $(\mathbf{r},t)$, for infinitesimally small δt. In other words, the equations (4)–(5) must be satisfied *locally* at point $(\mathbf{r},t)$ with a local and instantaneous velocity $\mathbf{v}(\mathbf{r},t)$:

$$-\partial_t \mathbf{E}(\mathbf{r},t) = \mathsf{D}\mathbf{E}(\mathbf{r},t)\mathbf{v}(\mathbf{r},t) \tag{15}$$
$$-\partial_t \mathbf{B}(\mathbf{r},t) = \mathsf{D}\mathbf{B}(\mathbf{r},t)\mathbf{v}(\mathbf{r},t) \tag{16}$$

In other words, if the RP, as it is believed, applies to all situations in electrodynamics, there must exist a local instantaneous velocity field $\mathbf{v}(\mathbf{r},t)$ satisfying (15)–(16) for all possible solutions of the following system of Maxwell–Lorentz equations:

$$\nabla\cdot\mathbf{E}\left(\mathbf{r},t\right) = \sum_{i=1}^{n} q^i \delta\left(\mathbf{r}-\mathbf{r}^i\left(t\right)\right) \tag{17}$$
$$c^2\nabla\times\mathbf{B}\left(\mathbf{r},t\right)-\partial_t\mathbf{E}\left(\mathbf{r},t\right) = \sum_{i=1}^{n} q^i \delta\left(\mathbf{r}-\mathbf{r}^i\left(t\right)\right)\mathbf{v}^i\left(t\right) \tag{18}$$
$$\nabla\cdot\mathbf{B}\left(\mathbf{r},t\right) = 0 \tag{19}$$
$$\nabla\times\mathbf{E}\left(\mathbf{r},t\right)+\partial_t\mathbf{B}\left(\mathbf{r},t\right) = 0 \tag{20}$$
$$m^i\gamma\left(\mathbf{v}^i\left(t\right)\right)\mathbf{a}^i(t) = q^i\left\{\mathbf{E}\left(\mathbf{r}^i\left(t\right),t\right)+\mathbf{v}^i\left(t\right)\times\mathbf{B}\left(\mathbf{r}^i\left(t\right),t\right) -c^{-2}\mathbf{v}^i\left(t\right)\left(\mathbf{v}^i\left(t\right)\cdot\mathbf{E}\left(\mathbf{r}^i\left(t\right),t\right)\right)\right\} \tag{21}$$
$$(i=1,2,\ldots,n)$$

where, $\gamma(\ldots) = \left(1 - \frac{(\ldots)^2}{c^2}\right)^{-\frac{1}{2}}$, q^i is the electric charge and m^i is the rest mass of the i-th particle. That is, substituting an arbitrary solution[2] of (17)–(21) into (15)–(16), the overdetermined system of equations must have a solution for $\mathbf{v}(\mathbf{r}, t)$.

However, one encounters the following difficulty:

Theorem 1. *There is a dense subset of solutions of the coupled Maxwell–Lorentz equations (17)–(21) for which there cannot exist a local instantaneous velocity field* $\mathbf{v}(\mathbf{r}, t)$ *satisfying (15)–(16).*

Proof. The proof is almost trivial for a locus $(\mathbf{r}, t)$ where there is a charged point particle. However, in order to avoid the eventual difficulties concerning the physical interpretation, we are providing a proof for a point $(\mathbf{r}_*, t_*)$ where there is assumed no source at all.

Consider a solution $\left(\mathbf{r}^1(t), \mathbf{r}^2(t), \ldots, \mathbf{r}^n(t), \mathbf{E}(\mathbf{r}, t), \mathbf{B}(\mathbf{r}, t)\right)$ of the coupled Maxwell–Lorentz equations (17)–(21), which satisfies (15)–(16). At point $(\mathbf{r}_*, t_*)$, the following equations hold:

$$-\partial_t \mathbf{E}(\mathbf{r}_*, t_*) = \mathsf{D}\mathbf{E}(\mathbf{r}_*, t_*)\mathbf{v}(\mathbf{r}_*, t_*) \tag{22}$$

$$-\partial_t \mathbf{B}(\mathbf{r}_*, t_*) = \mathsf{D}\mathbf{B}(\mathbf{r}_*, t_*)\mathbf{v}(\mathbf{r}_*, t_*) \tag{23}$$

$$\partial_t \mathbf{E}(\mathbf{r}_*, t_*) = c^2 \nabla \times \mathbf{B}(\mathbf{r}_*, t_*) \tag{24}$$

$$-\partial_t \mathbf{B}(\mathbf{r}_*, t_*) = \nabla \times \mathbf{E}(\mathbf{r}_*, t_*) \tag{25}$$

$$\nabla \cdot \mathbf{E}(\mathbf{r}_*, t_*) = 0 \tag{26}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}_*, t_*) = 0 \tag{27}$$

Without loss of generality we can assume—at point $\mathbf{r}_*$ and time t_*—that operators $\mathsf{D}\mathbf{E}(\mathbf{r}_*, t_*)$ and $\mathsf{D}\mathbf{B}(\mathbf{r}_*, t_*)$ are invertible and $v_z(\mathbf{r}_*, t_*) \neq 0$.

Now, consider a 3×3 matrix J such that

$$J = \begin{pmatrix} \partial_x E_x(\mathbf{r}_*, t_*) & J_{xy} & J_{xz} \\ \partial_x E_y(\mathbf{r}_*, t_*) & \partial_y E_y(\mathbf{r}_*, t_*) & \partial_z E_y(\mathbf{r}_*, t_*) \\ \partial_x E_z(\mathbf{r}_*, t_*) & \partial_y E_z(\mathbf{r}_*, t_*) & \partial_z E_z(\mathbf{r}_*, t_*) \end{pmatrix} \tag{28}$$

with

$$J_{xy} = \partial_y E_x(\mathbf{r}_*, t_*) + \lambda \tag{29}$$

$$J_{xz} = \partial_z E_x(\mathbf{r}_*, t_*) - \lambda \frac{v_y(\mathbf{r}_*, t_*)}{v_z(\mathbf{r}_*, t_*)} \tag{30}$$

[2]Without entering into the details, it must be noted that the Maxwell–Lorentz equations (17)–(21), exactly in this form, have *no* solution. The reason is that the field is singular at precisely the points where the coupling happens: on the trajectories of the particles. The generally accepted answer to this problem is that the real source densities are some "smoothed out" Dirac deltas, determined by the physical laws of the internal worlds of the particles—which are, supposedly, outside of the scope of CED. With this explanation, for the sake of simplicity we leave the Dirac deltas in the equations. Since our considerations here focuses on the electromagnetic field, satisfying the four Maxwell equations, we must only assume that there is a coupled dynamics—approximately described by equations (17)–(21)—and that it constitutes an initial value problem. In fact, Theorem 1 could be stated in a weaker form, by leaving the concrete form and dynamics of the source densities unspecified.

by virtue of which

$$\begin{aligned} J_{xy}v_y(\mathbf{r}_*,t_*)+J_{xz}v_z(\mathbf{r}_*,t_*) &= v_y(\mathbf{r}_*,t_*)\partial_y E_x(\mathbf{r}_*,t_*) \\ &\quad +v_z(\mathbf{r}_*,t_*)\partial_z E_x(\mathbf{r}_*,t_*) \end{aligned} \tag{31}$$

Therefore, $J\mathbf{v}(\mathbf{r}_*,t_*)=\mathsf{D}\mathbf{E}(\mathbf{r}_*,t_*)\mathbf{v}(\mathbf{r}_*,t_*)$. There always exists a vector field $\mathbf{E}^{\#}_{\lambda}(\mathbf{r})$ such that its Jacobian matrix at point $\mathbf{r}_*$ is equal to J. Obviously, from (26) and (28), $\nabla\cdot\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*)=0$. Therefore, there exists a solution of the Maxwell–Lorentz equations, such that the electric and magnetic fields $\mathbf{E}_\lambda(\mathbf{r},t)$ and $\mathbf{B}_\lambda(\mathbf{r},t)$ satisfy the following conditions:[3]

$$\mathbf{E}_\lambda(\mathbf{r},t_*) = \mathbf{E}^{\#}_{\lambda}(\mathbf{r}) \tag{32}$$

$$\mathbf{B}_\lambda(\mathbf{r},t_*) = \mathbf{B}(\mathbf{r},t_*) \tag{33}$$

At $(\mathbf{r}_*,t_*)$, such a solution obviously satisfies the following equations:

$$\partial_t\mathbf{E}_\lambda(\mathbf{r}_*,t_*) = c^2\nabla\times\mathbf{B}(\mathbf{r}_*,t_*) \tag{34}$$

$$-\partial_t\mathbf{B}_\lambda(\mathbf{r}_*,t_*) = \nabla\times\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*) \tag{35}$$

therefore

$$\partial_t\mathbf{E}_\lambda(\mathbf{r}_*,t_*)=\partial_t\mathbf{E}(\mathbf{r}_*,t_*) \tag{36}$$

As a little reflection shows, if $\mathsf{D}\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*)$, that is J, happened to be not invertible, then one can choose a *smaller* λ such that $\mathsf{D}\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*)$ becomes invertible (due to the fact that $\mathsf{D}\mathbf{E}(\mathbf{r}_*,t_*)$ is invertible), and, at the same time,

$$\nabla\times\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*)\neq\nabla\times\mathbf{E}(\mathbf{r}_*,t_*) \tag{37}$$

Consequently, from (36) , (30) and (22) we have

$$-\partial_t\mathbf{E}_\lambda(\mathbf{r}_*,t_*)=\mathsf{D}\mathbf{E}_\lambda(\mathbf{r}_*,t_*)\mathbf{v}(\mathbf{r}_*,t_*)=\mathsf{D}\mathbf{E}^{\#}_{\lambda}(\mathbf{r}_*)\mathbf{v}(\mathbf{r}_*,t_*) \tag{38}$$

and $\mathbf{v}(\mathbf{r}_*,t_*)$ is uniquely determined by this equation. On the other hand, from (35) and (37) we have

$$-\partial_t\mathbf{B}_\lambda(\mathbf{r}_*,t_*)\neq\mathsf{D}\mathbf{B}_\lambda(\mathbf{r}_*,t_*)\mathbf{v}(\mathbf{r}_*,t_*)=\mathsf{D}\mathbf{B}(\mathbf{r}_*,t_*)\mathbf{v}(\mathbf{r}_*,t_*) \tag{39}$$

because $\mathsf{D}\mathbf{B}(\mathbf{r}_*,t_*)$ is invertible, too. That is, for $\mathbf{E}_\lambda(\mathbf{r},t)$ and $\mathbf{B}_\lambda(\mathbf{r},t)$ there is no local and instantaneous velocity at point $\mathbf{r}_*$ and time t_*.

At the same time, λ can be arbitrary small, and

$$\lim_{\lambda\to0}\mathbf{E}_\lambda(\mathbf{r},t) = \mathbf{E}(\mathbf{r},t) \tag{40}$$

$$\lim_{\lambda\to0}\mathbf{B}_\lambda(\mathbf{r},t) = \mathbf{B}(\mathbf{r},t) \tag{41}$$

Therefore solution $\left(\mathbf{r}^1_\lambda(t),\mathbf{r}^2_\lambda(t),\ldots,\mathbf{r}^n_\lambda(t),\mathbf{E}_\lambda(\mathbf{r},t),\mathbf{B}_\lambda(\mathbf{r},t)\right)$ can fall into an arbitrary small neighborhood of $\left(\mathbf{r}^1(t),\mathbf{r}^2(t),\ldots,\mathbf{r}^n(t),\mathbf{E}(\mathbf{r},t),\mathbf{B}(\mathbf{r},t)\right)$.[4] □

[3] $\mathbf{E}^{\#}_{\lambda}(\mathbf{r})$ and $\mathbf{B}_\lambda(\mathbf{r},t_*)$ can be regarded as the initial configurations at time t_*; we do not need to specify a particular choice of initial values for the sources.

[4] Notice that our investigation has been concerned with the general laws of Maxwell–Lorentz electrodynamics of a coupled particles + electromagnetic field system. The proof of the

Thus, the meaning of the concept of "electromagnetic field moving with a local instantaneous velocity $\mathbf{v}(\mathbf{r}, t)$ at point $\mathbf{r}$ and time t", that we obtained by a straightforward generalization of the example of the stationary field of a uniformly moving charge, is untenable. We do not see other available rational meaning of this concept. Such a concept, on the other hand, would be a necessary conceptual plugin to the RP. In any event, lacking a better suggestion, we must conclude that the RP is a statement which is meaningless for a general electrodynamic situation.

4 No Persistence without Motion

There is a long debate in contemporary metaphysics whether and in what sense instantaneous velocity can be regarded as an intrinsic property of an object at a given moment of time (Butterfield 2006; Arntzenius 2000; Tooley 1988; Hawley 2001, 76–80; Sider 2001, 34–35). There seems to be, however, a consensus that

> [...] the notion of velocity presupposes the persistence of the object concerned. For average velocity is a quotient, whose numerator must be the distance traversed by the given persisting object: otherwise you could give me a superluminal velocity by dividing the distance between me and the Sun by a time less than eight minutes. So presumably, average velocity's limit, instantaneous velocity, also presupposes persistence. (Butterfield 2005, 257)

In this section we argue that the opposite is also true: the notion of persistence requires the existence of instantaneous velocity.

It is common to all theories of persistence—endurantism vs. perdurantism—that a persisting entity needs to have some package of individuating properties, in terms of which one can express that two things in two different spatiotemporal regions are identical, or at least constitute different spatiotemporal parts of the same entity. Butterfield writes:

> I believe that [the criteria of identity] are largely independent of the endurantism–perdurantism debate; and in particular, that endurantism and perdurantism [...] face some common questions about criteria of identity, and can often give the same, or similar, answers to them. [...] [A]ll parties need to provide criteria of identity for objects, presumably invoking the usual notions of qualitative similarity and-or causation (Butterfield 2005, 248–289)

Without loss of generality we may assume that each of these individuating properties can be characterized as such that a certain quantity f_i takes a certain value. Consider a primitive example: the redness of the ball in Fig. 2 can

theorem was essentially based on the presumption that all solutions of the Maxwell–Lorentz equations, determined by *any* initial state of the particles + electromagnetic field system, corresponded to physically possible configurations of the electromagnetic field. It is sometimes claimed, however, that the solutions must be restricted by the so called retardation condition, according to which all physically admissible field configurations must be generated from the retarded potentials belonging to some pre-histories of the charged particles (Jánossy 1971, p. 171; Frisch 2005, p. 145). There is no obvious answer to the question of how Theorem 1 is altered under such additional condition.

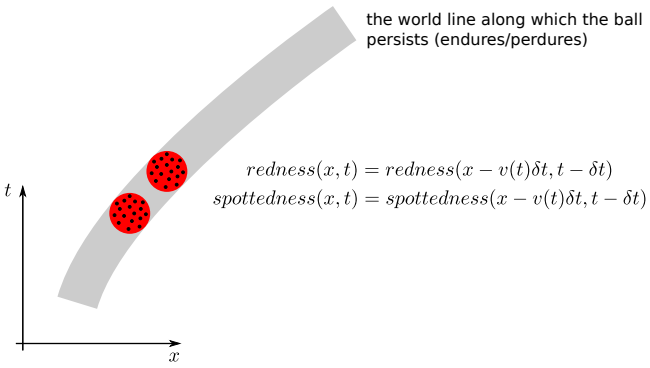


Figure 2: A ball is individuated by its redness, spottedness, etc.

be characterized as such that the wavelength of light reflected from the instantaneous surface of the ball is around 650 nm. Or, more abstractly, just imagine a quantity the spatiotemporal distribution of which takes value 1 in a region where redness is instantiated—for example, on the locus of the ball—and takes value 0 otherwise.

Now, in order to express the fact of persistence, consider a given n-tuple of individuating quantities $\{f_i\}_{i=1}^n$ that is supposed to trace out the trajectory or spacetime tube along which the entity persists. The different theories of persistence disagree in the actual content of the package $\{f_i\}_{i=1}^n$, these differences are not important from the point of view of our present concern. The following necessary condition is however common to all intuitions:

$$\begin{aligned} f_i(\mathbf{r},t) &= f_i(\mathbf{r} - \mathbf{v}\,(t)\,\delta t, t - \delta t) \\ &\qquad (i = 1, 2, \ldots, n) \end{aligned} \tag{42}$$

for all $(\mathbf{r}, t)$ where the object is present, at least for a small, infinitesimal, interval of time δt (Fig. 2), with some instantaneous velocity $\mathbf{v}\,(t)$. Without loss of generality we may assume that all functions in $\{f_i\}_{i=1}^n$ are smooth (if not, we can approximate them by smooth functions). Expressing (42) in a differential form, we have[5]

$$\begin{aligned} -\partial_t f_i(\mathbf{r},t) &= \mathsf{D} f_i(\mathbf{r},t)\mathbf{v}(t) \\ &\qquad (i = 1, 2, \ldots, n) \end{aligned} \tag{43}$$

In other words, the entity is *in motion* with some instantaneous velocities $\mathbf{v}(t)$. Let us call (43) the *equations of persistence*.

So far we considered the situation when the persistence can be formulated in terms of individuating quantities $\{f_i\}_{i=1}^n$ characterizing the entity in question *as a whole*. Generally, however, this is not necessarily the case. An extended object may persist, even if its holistic properties do not satisfy equations (43). Following however the same intuition by which we formulated the Mereological Principle of Motion, we propose the following thesis:

[5]For the sake of simplicity we may assume that all f_i are scalar functions, and $\mathsf{D} f_i$ is simply $\mathsf{grad}\, f_i$.

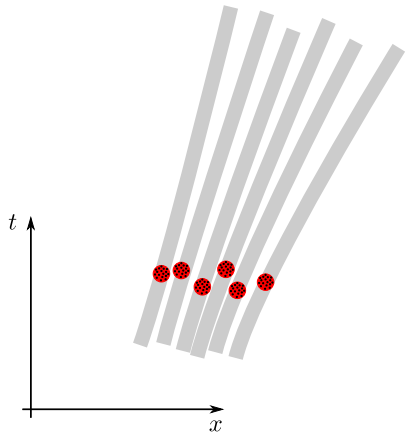


Figure 3: Persistence of an extended object requires the persitence of its local parts

Mereological Principle of Persistence (MPP) *If an extended object, as a whole, persists, then its all local parts persist.*

Accordingly, the persistence of an extended object requires the following condition for the spatial distributions:

$$f_i(\mathbf{r},t) = f_i(\mathbf{r}-\mathbf{v}\,(\mathbf{r},t)\,\delta t, t-\delta t) \quad (i=1,2,\ldots,n) \tag{44}$$

or

$$-\partial_t f_i(\mathbf{r},t) = \mathsf{D}f_i(\mathbf{r},t)\mathbf{v}(\mathbf{r},t) \quad (i=1,2,\ldots,n) \tag{45}$$

for all $(\mathbf{r},t)$ where the extended object is present; where $\mathbf{v}(\mathbf{r},t)$ is a local and instantaneous velocity field characterizing the *motion* of the local part of the extended entity at the spatiotemporal locus $(\mathbf{r},t)$ (Fig 3). Let us call (45) the *equations of persistence for an extended object.*

5 The Ontological Incompleteness of CED

As we have seen in Theorem 1, the distributions of the two fundamental electrodynamic field strengths, $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$, do not satisfy the equations of persistence (45). Therefore, the electromagnetic field individuated by the field strengths cannot be regarded as a persisting physical object; in other words, electromagnetic field cannot be regarded as being a real physical entity existing in space and time. This seems to contradict to the usual realistic interpretation of CED.

If electromagnetic field is a real entity persisting in space and time, then it cannot be individuated by the field strengths. That is, there must exist some

quantities other than the field strengths, perhaps outside of the scope of CED, individuating the local parts of electromagnetic field. This suggests that CED is an ontologically incomplete theory.

How to conceive properties, different from the field strengths, which are capable of individuating the electromagnetic field? One might think of them as some "finer", more fundamental, properties of the field, not only individuating it as a persisting extended object, but also determining the values of the field strengths. However, the following easily verifiable theorem shows that this determination cannot be so simple:

Theorem 2. *Let* $\{f_i\}_{i=1}^n$ *be a package of quantities for which there exist a local instantaneous velocity field* $\mathbf{v}(\mathbf{r},t)$ *satisfying the equations of persistence* (45) *in a given spacetime region. If a quantity* Φ *is a function of the quantities* $f_1, f_2, \ldots, f_n$ *in the following form:*

$$\Phi(\mathbf{r},t) = \Phi\left(f_1(\mathbf{r},t), f_2(\mathbf{r},t), ..., f_n(\mathbf{r},t)\right)$$

then Φ *also obeys the equation of persistence*

$$-\partial_t \Phi(\mathbf{r},t) \quad = \quad \mathsf{D}\Phi(\mathbf{r},t)\mathbf{v}(\mathbf{r},t)$$

with the same local instantaneous velocity field $\mathbf{v}(\mathbf{r},t)$*, within the same spacetime region.*

Therefore, $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$ cannot supervene pointwise upon some more fundamental individuating quantities satisfying the persistence equations. However, they might supervene in some non-local sense. For example, imagine that $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$ provide only a course-grained characterization of the field, but there exist some more fundamental fields $\mathbf{e}(\mathbf{r},t)$ and $\mathbf{b}(\mathbf{r},t)$, such that

$$\mathbf{E}(\mathbf{r},t) = \int_\Omega \mathbf{e}\left(\mathbf{r}',t'\right) d^4(\mathbf{r},t)$$
$$\mathbf{B}(\mathbf{r},t) = \int_\Omega \mathbf{b}\left(\mathbf{r}',t\right) d^4(\mathbf{r},t)$$

where Ω is a neighbourhood of $(\mathbf{r},t)$ (Fig. 4). In this case, the more fundamental quantities $\mathbf{e}(\mathbf{r},t)$ and $\mathbf{b}(\mathbf{r},t)$ may satisfy the equations of persistence, while $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$, supervening on $\mathbf{e}(\mathbf{r},t)$ and $\mathbf{b}(\mathbf{r},t)$, may not.

Acknowledgment

The research was partly supported by the OTKA Foundation, No. K100715.

References

Arntzenius, Frank. 2000. "Are There Really Instantaneous Velocities?" *The Monist* 83:187–208.

Brown, Harvey R. 2005. *Physical Relativity – Space-Time Structure from a Dynamical Perspectiv*e. Oxford, NY: Oxford University Press.

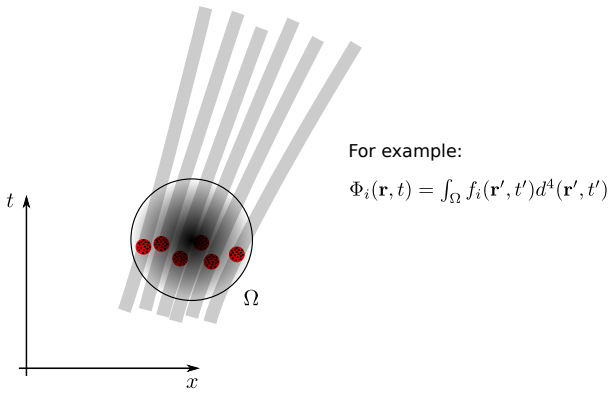


Figure 4: A non-local form of supervenience

Butterfield, Jeremy. 2005. "On the Persistence of Particles", *Foundations of Physics* 35:233–269.

Butterfield, Jeremy. 2006. "The Rotating Discs Argument Defeated", *British Journal for the Philosophy of Science* 57:1–45.

Frisch, Mathias. 2005. *Inconsistency, Asymmetry, and Non-Locality*, Oxford: Oxford University Press.

Galilei, Galileo. 1953. *Dialogue concerning the Two Chief World Systems, Ptolemaic & Copernican*, Berkeley: University of California Press.

Gömöri, Márton, and László E. Szabó. 2013. "Formal Statement of the Special Principle of Relativity", *Synthese*, DOI: 10.1007/s11229-013-0374-1

Hawley, Katherine. 2001. *How Things Persist*, Oxford: Oxford University Press.

Jackson, John David. 1999. *Classical Electrodynamics (Third edition).* Hoboken, NJ: John Wiley & Sons.

Jánossy, Lajos. 1971. *Theory of Relativity Based On Physical Reality*, Budapest: Akadémiai Kiadó.

Sider, Theodore. 2001. *Four-Dimensionalism*, Oxford: Oxford University Press.

Tooley, Michael. 1988. "In Defence of the Existence of States of Motion", *Philosophical Topics* 16:225–254.

A New Solution to the Problem of Old Evidence

Stephan Hartmann*

July 20, 2014

Abstract

The Problem of Old Evidence has troubled Bayesians ever since Clark Glymour first presented it in 1980. Several solutions have been proposed, but all of them have drawbacks and none of them is considered to be the definite solution. In this article, I propose a new solution which combines several old ideas with a new one. It circumvents the crucial omniscience problem in an elegant way and leads to a considerable confirmation of the hypothesis in question.

1 Introduction

The Problem of Old Evidence is easy to state. If the probability of the evidence, $P(\mathrm{E})$, is 1, then the likelihood $P(\mathrm{E}|\mathrm{H})$ is also 1, and hence

$$P^*(\mathrm{H}) := P(\mathrm{H}|\mathrm{E}) = \frac{P(\mathrm{E}|\mathrm{H})\,P(\mathrm{H})}{P(\mathrm{E})} = P(\mathrm{H})\,. \tag{1}$$

Consequently E does not confirm H according to standard Bayesian Confirmation Theory, i.e. if conditionalization is used to compute the posterior probability. This observation conflicts with the practice of science, as Glymour (1980) has forcefully pointed out.[1] Note that using conditionalization here is somewhat dubious as, in this case, nothing new is learned. So why should one even conditionalize?
In this article, I argue that something new happens in the course of the deliberation. The basic idea is this: Once a scientist becomes aware of the logical fact that the hypothesis under consideration entails the evidence (and that other available

*Munich Center for Mathematical Philosophy, LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich (Germany) – http://www.stephanhartmann.org – s.hartmann@lmu.de.

[1]See Earman (1992: ch. 5) for a critical discussion of a number of well-known responses to the Problem of Old Evidence.

theories do not entail E), she changes her belief about the disjunctive proposition A: "There is an alternative theory that entails the evidence or the evidence is the result of a chance mechanism." This, in turn, prompts an increase of the posterior probability of the hypothesis. Note that we are assuming that the scientist has *beliefs about the origin of the evidence*: E is either a deductive consequence of H, or it is a deductive consequence of an alternative to H, or it is the result of a chance mechanism, where the first and second disjunct are not mutually exclusive. Working out this idea in detail requires six elements:

(i) The scientist's beliefs are always consistent with the claim that she knows that the hypothesis H entails the evidence E. (This circumvents the crucial omniscience problem.) Initially, however, she is not aware of this logical fact.

(ii) Only after doing the deductions, the scientist becomes aware of the logical fact that H entails E (and that other available theories do not entail E). As a result of this,

(iii) the scientist lowers the probability of the disjunctive proposition A. Before the scientist becomes aware of the logical fact that H entails E, the scientist considered it to be quite likely that H does not entail E.[2] Instead she considers it to be quite likely that a (so far unknown) alternative to H entails E (and for which E is evidence) or that E is the result of a chance mechanism. Note that after becoming aware of the fact that H entails E (and that other available theories do not entail E), it is still possible that an alternative to H also entails E. In fact, E could be entailed by any given number of theories.

(iv) The probability of the evidence is $1 - \epsilon$, and not 1 (Fitelson 2004). This reflects the fact that E is a contingent proposition. Note, though, that ϵ can be arbitrarily small, and so this is not a strong restriction for all practical purposes. In fact we will calculate the limit $\epsilon \to 0$ at the end and it will turn out to be finite.

(v) The probability of the evidence does not change in the course of deliberation. More specifically, it is not affected by our deliberations about H or by the possible existence of an alternative to H. This poses a constraint on the likelihoods before and after becoming aware of the logical fact that H entails E (and that other available theories do not entail E).

[2]One could also argue that the scientist should be *agnostic* about the deductive relationship between H and E before she becomes aware of the logical fact that H entails E. We do not consider this possibility here because there is no agreement on how to model ignorance in a Bayesian framework. Cf. Norton (2008, 2011).

(vi) The scientist determines the posterior probability distribution by minimizing the Kullback-Leibler divergence between the posterior probability distribution and the prior probability distribution. Note that this procedure is more general than conditionalization: It leads to standard conditonalization and Jeffrey conditionalization if corresponding constraints concerning the posterior probability of the evidence are added (see Diaconis and Zabell 1982). More recently, Hartmann and Rafiee Rad (2014) have argued that minimizing the Kullback-Leibler divergence can also be used to model the learning of an indicative conditional and that standard objections or problems such as the Judy Benjamin Problem can be rebutted provided that the causal structure of the problem at hand is properly taken into account. Encouraged by these success stories, we conjecture that minimizing the Kullback-Leibler divergence between the posterior and the prior probability distribution is an essential ingredient in any general Bayesian account of belief change.

Note that on our proposal one does not have to conditionalize on E (which arguably does not make sense as the scientist does not learn E) or on the new proposition (after extending the language appropriately) that the hypothesis entails the evidence (Garber 1983). It also avoids counterfactual reasoning as in Howson's solution to the Problem of Old Evidence (Howson 1991).
The remainder of this article is organized as follows. Section 2 presents our general model which works for all cases where the probability of the evidence does not change in the course of deliberation. Section 3 then considers the special case of old evidence where the probability of the evidence is (close to) 1. We conclude, in Section 4, with a short remark concerning the adequacy of the proposed solution.

2 The General Model

We introduce the two usual binary propositional variables.[3] The variable H has the values H: "The hypothesis is true", and ¬H: "The hypothesis is false". The variable E has the values E: "The evidence obtains", and ¬E: "The evidence does not obtain". We consider the case where it is a matter of fact that H entails E. Additionally, we introduce the binary propositional variable A_T which has the values $\mathrm{A_T}$: "There is an alternative hypothesis that entails E", and $\neg\mathrm{A_T}$: "There is no alternative hypothesis that entails E" and the binary propositional variable A_C which has the values $\mathrm{A_C}$: "E is the result of a chance mechanism", and $\neg\mathrm{A_C}$:

[3]Throughout this article we follow the convention, adopted e.g. in Bovens and Hartmann (2003), that propositional variables are printed in (upper case) italic script, and that the instantiations of these variables are printed in (upper case) roman script. Bovens and Hartmann (2003) also introduce the bits of the theory of Bayesian Networks that we use in this article.

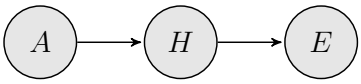


Figure 1: The Bayesian Network for the Problem of Old Evidence.

"E is not the result of a chance mechanism". Let A:= $\mathrm{A_T} \vee \mathrm{A_C}$. We assume that the scientist has beliefs about A, E and H and that it is a logical fact that H entails E.

The Bayesian Network in Figure 1 encodes the probabilistic dependencies and independencies between the three variables: The variable A directly influences the variable H, the variable H directly influences the variable E, and A influences E only through H: If the hypothesis is true (false), then our credence in E should be equal to the likelihood $P(\mathrm{E}|\mathrm{H})$ (or $P(\mathrm{E}|\neg\mathrm{H})$, respectively).

To complete the Bayesian Network, we have to fix the prior probability of the root node A and the conditional probabilities of all other nodes, given the values of their parents. First, we set

$$P(\mathrm{A}) = a \tag{2}$$

and assume that $a \in (0, 1)$. The value of a will, in fact, be fairly large as the scientist is not yet aware of the logical fact that H entails E. And so she strongly believes that there is an alternative to H that entails E or that E is the result of a chance mechanism.

Second, we set

$$P(\mathrm{H}|\mathrm{A}) = \alpha \quad , \quad P(\mathrm{H}|\neg\mathrm{A}) = 1\,, \tag{3}$$

with $\alpha \in (0, 1)$. If A is false, i.e. if there is no alternative to H that accounts for E and if E is not the result of a chance mechanism, then the probability of H is one. However, if $\mathrm{A_T}$ (and therefore A) is true and there is an alternative to H that entails E, then it is possible that H is true and entails E as well because there may be several theories that entail E. Hence $P(\mathrm{H}|\mathrm{A}) =: \alpha > 0$, where α measures how strongly the scientist believes in H, even if there is an alternative theory that entails E. It is important to note that $\alpha > 0$ implies that the propositions A and H are not mutually exclusive.

From eqs. (2) and (3) we calculate the prior probability of the hypothesis using the Law of Total Probability:

$$\begin{aligned} P(\mathrm{H}) &= P(\mathrm{H}|\mathrm{A})\,P(\mathrm{A}) + P(\mathrm{H}|\neg\mathrm{A})\,P(\neg\mathrm{A}) \\ &= \alpha\, a + \overline{a} \end{aligned} \tag{4}$$

From this equation it follows that α is fairly small as a is fairly large and $P(\mathrm{H})$ is

not very large. Typical values could be $a = .8$ and $\alpha = .2$, in which case we get $P(\mathrm{H}) = .36$, but our argument of course does not depend on these assignments. Finally, we set

$$P(\mathrm{E}|\mathrm{H}) = 1 \quad , \quad P(\mathrm{E}|\neg\mathrm{H}) = q\,, \tag{5}$$

with $q \in (0, 1)$. Here we assume that the beliefs of the scientist are consistent with the logical fact that H entails E.[4] She therefore assigns the conditional probability $P(\mathrm{E}|\mathrm{H})$ the value 1 which makes sure that her beliefs are coherent.[5] Setting $P(\mathrm{E}|\mathrm{H}) = 1$ may be ad hoc for $P(\mathrm{E}) \ll 1$ as the scientist is not yet aware of the logical fact that H entails E. However, if $P(\mathrm{E})$ is close to 1 (which is the case in the Problem of Old Evidence), then setting $P(\mathrm{E}|\mathrm{H})$ to 1 (or to a value close to 1) is a consequence of having coherent beliefs as we will show in Section 3.
The prior probability distribution over A, H and E is then given by

$$\begin{aligned} P(\mathrm{A},\mathrm{H},\mathrm{E}) = \alpha\, a \qquad &, \qquad P(\mathrm{A},\neg\mathrm{H},\mathrm{E}) = \overline{\alpha}\, a\, q \\ P(\mathrm{A},\neg\mathrm{H},\neg\mathrm{E}) = \overline{\alpha}\, a\, \overline{q} \qquad &, \qquad P(\neg\mathrm{A},\mathrm{H},\mathrm{E}) = \overline{a}\,. \end{aligned} \tag{6}$$

For all other instantiations of A, H and E, the prior probability vanishes. Here we have used the convenient shorthand $\overline{x} := 1 - x$, which we will use throughout this article. Here and in the remainder we also use the shorthand notation $P(\mathrm{A},\mathrm{H},\mathrm{E})$ for $P(\mathrm{A} \wedge \mathrm{H} \wedge \mathrm{E})$.
With this, we calculate

$$\begin{aligned} P(\mathrm{E}) &= \alpha\, a + \overline{\alpha}\, a\, q + \overline{a} \\ &= 1 - \overline{\alpha}\, a\, \overline{q}\,. \end{aligned} \tag{7}$$

Next, the scientist becomes aware of the logical fact that H entails E (and that other available theories do not entail E), which prompts her to change her beliefs. More specifically, she reduces the probability of A and sets

$$P'(\mathrm{A}) = a' < a, \tag{8}$$

where P' is the posterior probability measure. Given that the scientist is now aware of the logical fact that H entails E, she does not consider it so probable anymore that there is an alternative theory that entails E or that E is the result of a chance mechanism. Before becoming aware of the logical fact that H entails E, she strongly believed that there is an alternative to H that entails E or that

[4]Note that it is perfectly fine to set $P(\mathrm{E}|\mathrm{H}) = 1$ even if $P(\mathrm{E}) = 1 - \epsilon$. We might be uncertain about whether E and/or H obtain, but we may nevertheless be convinced that E is always true if H is true.

[5]Note that we are only considering deterministic theories here. For indeterministic theories, $P(\mathrm{E}|\mathrm{H}) < 1$ and our model does not apply in the present form. We leave the confirmation-theoretical analysis of old evidence for an indeterministic theory for another occasion.

E is the result of a chance mechanism (although her degrees of belief have always been consistent with the logical fact that H entails E). After becoming aware of the logical fact that H entails E (and that other available theories do not entail E), she still deems it possible that there is an alternative to H that entails E, but she does not believe in it so strongly anymore.[6]

Note that the belief change expressed in eq. (8) does not result from conditionalization. It is more similar to what actually happens in (Jeffrey) conditionalization when the probability of the evidence suddenly changes from one value to another. In the case of (Jeffrey) conditionalization, this change is prompted by an experience. In the case considered in this article, it is prompted by the insight that H entails E (and that other available theories do not entail E).

As H entails E, the scientist also sets

$$P'(\mathrm{E}|\mathrm{H}) = 1\,, \tag{9}$$

which is in line with the corresponding assignment in the prior probability distribution (see eq. (5)).

To proceed, we assume that the Bayesian Network depicted in Figure 1 remains unchanged after the agent changed her beliefs about A. Hence, the posterior probability distribution has the following form:

$$\begin{aligned} P'(\mathrm{A},\mathrm{H},\mathrm{E}) = \alpha' a' \quad &, \quad P'(\mathrm{A},\neg\mathrm{H},\mathrm{E}) = \overline{\alpha'}\, a'\, q' \\ P'(\mathrm{A},\neg\mathrm{H},\neg\mathrm{E}) = \overline{\alpha'}\, a'\, \overline{q'} \quad &, \quad P'(\neg\mathrm{A},\mathrm{H},\mathrm{E}) = \overline{a'}\,, \end{aligned} \tag{10}$$

where we have replaced all variables with the corresponding primed variables. Note that the value of a' is already fixed (as the scientist sets it to a lower value after becoming aware of the logical fact that H entails E and that other available theories do not entail E)[7], but the values of α' and q' have to be determined. We do this by minimizing the Kullback-Leibler divergence between P' and P taking all relevant constraints on P' into account.

Here are two additional constraints on the posterior probability distribution (besides eq. (8)). First, we assume that the probability of E remains unchanged in the course of deliberation, i.e.

$$P'(\mathrm{E}) = 1 - \overline{\alpha'}\, a'\, \overline{q'} \equiv P(\mathrm{E}) = 1 - \overline{\alpha}\, a\, \overline{q}\,. \tag{11}$$

[6]It could be objected that a' should be greater than a because it is easy to construct alternatives to H that also entail E after becoming aware of the logical fact that H entails E. Just add an irrelevant conjunct to H or make a Goodman-style move. If one does so, then the probability of H goes down. Hence, a scientist who has coherent beliefs and who reasons in this way will hold the view that H is not confirmed.

[7]Note, though, that a' has to satisfy inequality (13).

Hence,

$$\overline{\alpha}\, a\, \overline{q} - \overline{\alpha'}\, a'\, \overline{q'} = 0\,. \tag{12}$$

Next, we conclude from eq. (12) that $\overline{\alpha}\, a\, \overline{q} = \overline{\alpha'}\, a'\, \overline{q'} < a'$. Hence, we obtain

$$a' > \overline{\alpha}\, a\, \overline{q}\,, \tag{13}$$

as our third and final constraint on P'. Inequality (13) tells us that the scientist cannot reduce a' to an arbitrarily low value. A rational agent should always take the possible existence of an alternative theory that entails E or the possibility that E resulted from a chance mechanism into account.
In this section we will only assume that the probability of E does not change after the scientist becomes aware of the logical fact that H entails E (and that other available theories do not entail E) and after she lowered the probability of A in turn. In the next section, we will focus on the specific case that $P(\mathrm{E}) = P'(\mathrm{E})$ is close to 1, which is the situation in the Problem of Old Evidence.
We can now show the following theorem (proof in the Appendix).

Theorem: *Consider the Bayesian Network in Figure 1 with the prior probability distribution P from eq. (6). We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the posterior distribution P' is constrained by eqs. (8), (12) and (13), and (iii) P' minimizes the Kullback-Leibler divergence between P' and P. Let $\Delta := P'(\mathrm{H}) - P(\mathrm{H})$. Then*

$$\Delta = \frac{q}{\alpha + \overline{\alpha}\, q} \cdot \overline{\alpha}\, (a - a') \tag{14}$$

Hence, $\Delta > 0$ if and only if $a' < a$. And so we conclude that H is confirmed in the considered situation although the probability of the evidence does not change in the course of deliberation. Note that H is not directly confirmed by E. The confirmation is *indirect* as we first become aware of the logical fact that H entails E (and that other available theories do not entail E) and then, in turn, reduce the probability of A. This, then, results in an increase in the probability of H if we determine the posterior probability distribution by minimizing the Kullback-Leibler divergence between the posterior probability distribution and the prior probability distribution.

3 The Problem of Old Evidence

Let us now turn to the Problem of Old Evidence. So far we have only assumed that the probability of E does not change in the course of deliberation. In the

old-evidence situation, more specifically, the probability of E is close to 1. So let us set

$$P(\mathrm{E}) = 1 - \epsilon \tag{15}$$

with $\epsilon \ll 1$. Next, we observe from eq. (7) that

$$\epsilon = \overline{\alpha}\, a\, \overline{q}\,, \tag{16}$$

which has to be small. To achieve this, we set

$$q = 1 - \eta \tag{17}$$

with $\eta \ll 1$. Hence, $\epsilon = \overline{\alpha}\, \eta\, a < \eta \ll 1$. We do not have to impose any further constraints on the values of a and α. Note that the choice (17) makes a lot of sense: The scientist knows that $P(\mathrm{E})$ is close to 1, independently of whether H is true or false: It is simply a contingent fact that E obtains (note that ϵ and η can be set to an arbitrarily small value and so our proposed solution of the Problem of Old Evidence works, as we will see, for all practical purposes). However, the scientist does not know whether or not H is true, and so she should set $P(\mathrm{E}|\mathrm{H})$ and $P(\mathrm{E}|\neg\mathrm{H})$ to 1 or to a value very close to 1 (again, there is no difference between the two assignments for all practical purposes). It is important to note that setting $P(\mathrm{E}|\mathrm{H})$ and $P(\mathrm{E}|\neg\mathrm{H})$ to 1 (or to a value close to 1) does not mean that the scientist knows that E is a deductive consequence of H: In the present case, the scientist has to make these assignments simply in order to have coherent beliefs.[8]

To proceed, let us plot Δ as a function of α for plausible values of a, a' and q. Figure 2 shows that one gets a considerable amount of confirmation (for $\alpha = .2$, for example, the probability rises by .15). It is also plausible, as Figure 2 suggests, that one gets more confirmation for smaller values of α: if we strongly believe that there is an alternative to H that entails E or that E is the result of a chance mechanism and if we also strongly believe that H is false if there is an alternative to H or if E is the result of a chance mechanism, then we will be quite impressed by becoming aware of the logical fact that H entails E. Hence, *ceteris paribus*, smaller values of α lead to more confirmation of H than larger values of α.

All this can also be seen analytically. To do so, we expand Δ in a Taylor-series up to zeroth order in η and obtain:

$$\Delta = \overline{\alpha}\,(a - a') + \mathcal{O}(\eta) \tag{18}$$

[8]Note that $P(\mathrm{E}) = 1$ implies that $P(\mathrm{E}|\mathrm{H_i}) = 1$ for every element H_i of a partition H_1, H_2,... with a non-vanishing prior probability. However, from $P(\mathrm{E}) = 1 - \epsilon$ we cannot infer that $P(\mathrm{E}|\mathrm{H_i}) = 1 - \epsilon'$ with $\epsilon' \ll 1$. In the present case, however, the probability of H (as well as the probability of $\neg$H) is neither very small nor very large. And so the scientist has to set $P(\mathrm{E}|\mathrm{H})$ and $P(\mathrm{E}|\neg\mathrm{H})$ to 1 (or to a value close to 1) in order to have coherent beliefs.

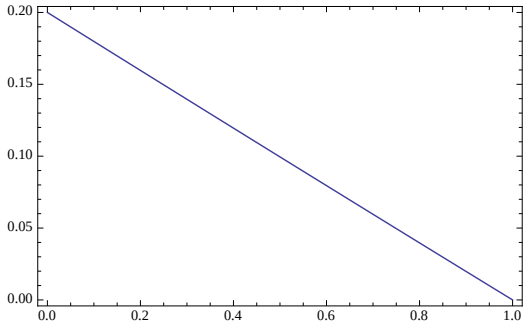


Figure 2: Δ as a function of α for $a = .8, a' = .6$ and $q = .99$.

Hence, H is more confirmed, the smaller the value of α (for fixed values of a and $a' > \overline{\alpha}\,\eta\,a$).
Note that we can take the limit $\eta \to 0$ from eq. (18) and obtain

$$\Delta = \overline{\alpha}\,(a - a'). \tag{19}$$

This result is approached as $P(\mathrm{E})$ approaches 1, so that our assumption that $P(\mathrm{E}) = 1 - \epsilon$ was only a mathematical trick that we had to make to proceed with the calculation. Interestingly, however, eq. (19) can also be obtained if one uses Jeffrey conditionalization in a straightforward way. To do so, we calculate the posterior probability of H after learning that the probability of A shifted:

$$\begin{aligned} P^*(\mathrm{H}) &= P(\mathrm{H}|\mathrm{A})\,P^*(\mathrm{A}) + P(\mathrm{H}|\neg\mathrm{A})\,P^*(\neg\mathrm{A}) \\ &= \alpha\,a' + \overline{a'}\,. \end{aligned} \tag{20}$$

With eq. (4), we obtain

$$\begin{aligned} \Delta^* &:= P^*(\mathrm{H}) - P(\mathrm{H}) \\ &= \overline{\alpha}\,(a - a')\,. \end{aligned} \tag{21}$$

Note that this calculation assumes that α is fixed and therefore does not change in the course of deliberation.[9] In hindsight, this assumption is justified as it turns out that α changes only slightly (details in the Appendix). It was, however, not at all clear from the beginning that this is the case, and so we were not justified to use Jeffrey conditionalization in the first place. Conditionalization and Jeffrey conditionalization often work, but they do not always work. The more general procedure to rationally change one's beliefs is to use the method of minimizing the Kullback-Leibler divergence taking into account the new information as a constraint on the posterior probability distribution.

[9]This can be seen by comparing eqs. (20) and (34).

In closing, let us note that our solution to the Problem of Old Evidence also shows that old evidence typically disconfirms the theory that preceded H. For example, the advanced perihelion of Mercury (= E) disconfirms Newtonian Mechanics (= H'). As $P(\mathrm{E}) = 1 - \epsilon$, we set (as before) $P(\mathrm{E}|\mathrm{H}')$ and $P(\mathrm{E}|\neg\mathrm{H}')$ to 1 or to a value close to 1. The scientist then becomes aware of the logical fact that H' does *not* entail E. She therefore increases the probability of A, i.e. she sets $a' > a$: It is now more probable that there is an alternative to H' that entails E or that E is the result of a chance mechanism. In fact, she has to set $a' = 1$ as she is now certain that there is either another theory that entails E or that E is the result of a chance mechanism. Hence, according to eq. (19), H' is disconfirmed and $P'(\mathrm{H}') < P(\mathrm{H}')$. We conjecture that it will be difficult to plausibly show this in Howson's counterfactual approach (Howson 1991).

4 Concluding Remark

Our argument crucially depends on the assumption that scientists believe the proposition A with a certain probability and that they change the corresponding probability once they become aware of the logical fact that H entails E (and that other available theories do not entail E). This is an empirical assumption, and it has be be investigated in detail whether it is true or false for concrete examples. We leave this for another occasion.

Acknowledgements

Thanks to Aidan Lyon for an eye-opening discussion. I would also like to thank Seamus Bradley, Erik Curiel, Igor Douven, Clark Glymour, Jim Joyce, Jan Sprenger, and Greg Wheeler for helpful feedback on an earlier draft.

A Appendix: Proof of the Theorem

Let P and P' be two probability distributions. The Kullback-Leibler divergence $D_{KL}(P'||P)$ between P' and P is defined as

$$D_{KL}(P'||P) := \sum_{i=1}^{n} P'(S_i) \log \frac{P'(S_i)}{P(S_i)} \,. \tag{22}$$

Here $S_1, \ldots, S_n$ be the possible values of a random variable S over which probability distributions P' and P are defined.

Using eqs. (6) and (10), we obtain:

$$\begin{aligned} D_{KL}(P'||P) &:= \sum_{A,H,E} P'(A,H,E)\log\frac{P'(A,H,E)}{P(A,H,E)} \\ &= \alpha'\,a'\log\frac{\alpha'\,a'}{\alpha\,a} + \overline{\alpha'}\,a'\,q'\log\frac{\overline{\alpha'}\,a'\,q'}{\overline{\alpha}\,a\,q} \qquad (23) \\ &\quad +\overline{\alpha'}\,a'\,\overline{q'}\log\frac{\overline{\alpha'}\,a'\,\overline{q'}}{\overline{\alpha}\,a\,\overline{q}} + \overline{a'}\log\frac{\overline{a'}}{\overline{a}} \\ &= a'\,\log\frac{a'}{a} + \overline{a'}\,\log\frac{\overline{a'}}{\overline{a}} + a'\left(\alpha'\,\log\frac{\alpha'}{\alpha} + \overline{\alpha'}\,\log\frac{\overline{\alpha'}}{\overline{\alpha}}\right) \qquad (24) \\ &\quad +\overline{\alpha'}\,a'\left(q'\,\log\frac{q'}{q} + \overline{q'}\,\log\frac{\overline{q'}}{\overline{q}}\right) \qquad (25) \end{aligned}$$

Next, we minimize

$$L := D_{KL}(P'||P) + \lambda(\overline{\alpha}\,a\,\overline{q} - \overline{\alpha'}\,a'\,\overline{q'}) \qquad (26)$$

with respect to α' and q'. Here λ is a Lagrange multiplier, and the expression in the bracket takes the constraint from eq. (12) into account.
To find the minimum, we first differentiate L with respect to q' and obtain:

$$\frac{\partial L}{\partial q'} = \overline{\alpha'}\,a'\left(\log\left(\frac{q'}{\overline{q'}}\,\frac{\overline{q}}{q}\right) + \lambda\right) \qquad (27)$$

Setting this expression equal to zero and noting that $a' > 0$ and $\alpha' < 1$, we obtain

$$q' = \frac{q}{q + \overline{q}\,\mathrm{e}^{\lambda}}\,. \qquad (28)$$

With this, L simplifies to

$$\begin{aligned} L &= a'\,\log\frac{a'}{a} + \overline{a'}\,\log\frac{\overline{a'}}{\overline{a}} + a'\left(\alpha'\,\log\frac{\alpha'}{\alpha} + \overline{\alpha'}\,\log\frac{\overline{\alpha'}}{\overline{\alpha}}\right) \\ &\quad -\overline{\alpha'}\,a'\log(q + \overline{q}\,\mathrm{e}^{\lambda}) + \overline{\alpha}\,\lambda\,a\,\overline{q} \qquad (29) \end{aligned}$$

Next, we differentiate this expression with respect to α' and obtain

$$\frac{\partial L}{\partial \alpha'} = a'\log\left(\frac{\alpha'}{\overline{\alpha'}}\,\frac{\overline{\alpha}}{\alpha}\,(q + \overline{q}\,\mathrm{e}^{\lambda})\right)\,. \qquad (30)$$

Setting this expression equal to zero and noting that $a' > 0$, we obtain

$$\alpha' = \frac{\alpha}{\alpha + \overline{\alpha}\,(q + \overline{q}\,\mathrm{e}^{\lambda})}\,. \qquad (31)$$

Inserting eqs. (28) and (31) into eq. (12), we obtain

$$\mathrm{e}^{\lambda} = \frac{(\alpha + \overline{\alpha}\, q)\, a}{a' - \overline{\alpha}\, a\, \overline{q}}\,. \tag{32}$$

Note that the denominator in eq. (32) is always greater than zero because of the constraint expressed in eq. (12).
Inserting eq. (32) into eqs. (28) and (31), after a short calculation we obtain

$$\alpha' = \frac{\alpha}{a'} \cdot \frac{a' - \overline{\alpha}\, a\, \overline{q}}{\alpha + \overline{\alpha}\, q} \quad , \quad q' = \frac{(a' - \overline{\alpha}\, a\, \overline{q})\, q}{a'\, q + \alpha\, a\, \overline{q}}\,. \tag{33}$$

This completes the calculation of the posterior probability distribution.
Let us now explore under which conditions the posterior probability of H, i.e. $P'(\mathrm{H})$ is greater than the prior probability of H, i.e. $P(\mathrm{H})$. That is, let us ask under which conditions H is confirmed. To do so, we calculate

$$P'(\mathrm{H}) = \alpha'\, a' + \overline{a'}\,. \tag{34}$$

Hence, using eq. (4), we finally obtain

$$\begin{aligned} \Delta &:= P'(\mathrm{H}) - P(\mathrm{H}) \\ &= \alpha'\, a' - \alpha\, a + (a - a') \\ &= \frac{q}{\alpha + \overline{\alpha}\, q} \cdot \overline{\alpha}\, (a - a')\,, \end{aligned} \tag{35}$$

which completes the proof of the theorem.

Let us finally consider α' and q' for the old evidence situation. We will show that $\alpha' \leq \alpha$ and $q' \leq q$ for $a > a'$. Setting $a' = \overline{\alpha}\, \eta\, a + \delta$ with $\delta > 0$ (see eq. (13)), we obtain

$$\begin{aligned} \frac{\alpha'}{\alpha} &= \frac{\delta}{(\overline{\alpha}\, \eta\, a + \delta)(\alpha + \overline{\alpha}\, \overline{\eta})} = \frac{\delta}{\delta + \overline{\alpha}\, \eta\, (a - \overline{\alpha}\, \eta\, a - \delta)} \\ &= \frac{\delta}{\delta + \overline{\alpha}\, \eta\, (a - a')} \leq 1\,. \end{aligned} \tag{36}$$

Similarly,

$$\begin{aligned} \frac{q'}{q} &= \frac{\delta}{(\overline{\alpha}\, \eta\, a + \delta)\, \overline{\eta} + \alpha\, \eta\, a} = \frac{\delta}{\delta + \eta\, (a - \overline{\alpha}\, \eta\, a - \delta)} \\ &= \frac{\delta}{\delta + \eta\, (a - a')} \leq 1\,. \end{aligned} \tag{37}$$

Hence, for $\eta \ll \delta$ and $a > a'$, we find that $\alpha' \lesssim \alpha$ and $q' \lesssim q$.

References

Bovens, L. and S. Hartmann (2003). *Bayesian Epistemology.* Oxford: Oxford University Press.

Diaconis, P. and S. Zabell (1982). Updating Subjective Probability, *Journal of the American Statistical Association* 77: 822–830.

Earman, J. (1992). *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory.* Cambridge: Bradford-MIT.

Fitelson, B. (2004). Earman on Old Evidence and Measures of Confirmation. Unpublished manuscript. University of California-Berkeley.

Garber, D. (1983). Old Evidence and Logical Omniscience in Bayesian Confirmation Theory. In: J. Earman (ed.): *Testing Scientific Theories.* Minneapolis: The University of Minnesota Press, pp. 99–131.

Glymour, C. (1980). Why I am not a Bayesian. In: *Theory and Evidence.* Princeton: Princeton University Press, pp. 63–93.

Hartmann, S. and S. Rafiee-Rad (2014). Learning Conditionals. MCMP Working Paper.

Howson, C. (1991). The Old Evidence Problem. *The British Journal for the Philosophy of Science* 42: 547–555.

Norton, J. D. (2008). Ignorance and Indifference. *Philosophy of Science* 75: 45–68.

Norton, J. D. (2011). Challenges to Bayesian Confirmation Theory. In: P. Bandyopadhyay and M. Forster (eds.): *Handbook of the Philosophy of Science, Vol. 7: Philosophy of Statistics.* Amsterdam: Elsevier, pp. 391–440.

A New Interpretation of the Representational Theory of Measurement

Conrad Heilmann*

Forthcoming in *Philosophy of Science*

Abstract

On the received view, the Representational Theory of Measurement reduces measurement to the numerical representation of empirical relations. This account of measurement has been widely criticised. In this paper, I provide a new interpretation of the Representational Theory of Measurement that sidesteps these debates. I propose to view the Representational Theory of Measurement as a library of theorems that investigate the numerical representability of qualitative relations. Such theorems are useful tools for concept formation which, in turn, is one crucial aspect of measurement for a broad range of cases in linguistics, rational choice, metaphysics, and the social sciences.

1 Introduction

The Representational Theory of Measurement (RTM) is one of the main accounts of measurement (Swistak, 1990; Boumans, 2008; Cartwright and Chang, 2008). It characterises measurement as a mapping between two relational structures, an empirical one and a numerical one (Krantz *et al.*, 1971; Suppes *et al.*, 1989; Luce *et al.*, 1990).

RTM is much criticised. Its critics, such as those that endorse a realist or operationalist conception of measurement, focus mainly on the fact that RTM advances an abstract conception of measurement that is not connected to empirical work as closely as it should be: it reduces measurement to representation, without specifying the actual process of measuring something, and problems like measurement error and the construction of reliable measurement instruments are ignored (Michell, 1990; Decoene *et al.*, 1995; Michell, 1995; Boumans, 2007; Reiss, 2008).

*Erasmus Institute for Philosophy and Economics (EIPE), Faculty of Philosophy, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, Email: heilmann@fwb.eur.nl

In this paper, I do not engage with these worries, but rather sidestep them by proposing to interpret RTM in a different way. I will not assess RTM as a candidate theory of measurement, but propose the following two-step interpretation: firstly, RTM should be viewed as simply providing a library of mathematical theorems. Secondly, RTM theorems have a particular structure that makes them useful for investigating problems of concept formation. More precisely, I propose to view theorems in RTM as providing us with mathematical structures which, if sustained by specific conceptual interpretations, can provide insights into the possibilities and limits of representing concepts numerically. If we adopt this interpretation, there is no reason why RTM theorems should be restricted to specifying the conditions under which only *empirical* relations can be represented numerically. Rather, we can view the theorems as providing insights into how to numerically represent any sort of qualitative relation between any sort of object. Indeed, those objects can include highly idealised or hypothetical ones. On this view, RTM is no longer viewed as candidate for a full-fledged theory of measurement, but rather as a tool that can be used in discussing the formation of concepts, which in turn is often a particularly difficult part of measurement, especially in the social sciences.

This new interpretation of RTM has a number of advantages. Firstly, it allows us to use RTM theorems in the investigation of abstract concepts. All this means is that since we move from an empirical relational structure to a more general qualitative relational structure, we can also ask what kind of qualitative relations between imagined or idealised objects could be represented numerically. This is helpful in areas of inquiry in which there no developed empirical concepts and where there is a lack agreement on a number of basic questions (such as cases in linguistics, rational choice, metaphysics, and the social sciences more generally). Secondly, the new interpretation gives more flexibility to engage in 'backwards engineering' of foundations for quantitative concepts. In contexts in which we operate with numbers that lack adequate conceptual and epistemic foundations, we can investigate what kinds of qualitative relations between what kinds of objects would need to exist in order to motivate the kind of numbers that are already in use. That is, we can look at areas of inquiry that use quantities that are not derived from a measurement process and investigate whether these quantities can be seen as numerical representations of qualitative relations (and, in turn, whether such a representation can help in devising a measurement process). Thirdly, the interpretation serves as a way to rehabilitate RTM as a useful part of theoretical tools for measurement.

I proceed as follows. Section 2 introduces RTM in its received interpretation. Section 3 explains the new interpretation of RTM as representing qualitative instead of only empirical relations. Section 4 discusses desiderata of the interpretation. Section 5 briefly concludes.

2 The Representational Theory of Measurement

Before RTM, measurement was mainly associated with the idea that (physical) quantities are assigned numbers (Russell, 1903, 176). RTM has taken a more abstract stance, substituting the idea of physical quantity or magnitude with properties or features of objects or with relations between such properties or features (Swistak, 1990). Swistak (1990, 7) also maintains that the 'representational paradigm is the fundamental notion of measurement which is in use in the contemporary theory of measurement' and ascribes the coining of the term 'representational theory of measurement' to Adams (1966). The authoritative statement of RTM can be found in the three books Krantz *et al.* (1971), Suppes *et al.* (1989) and Luce *et al.* (1990), which are building on earlier axiomatic work by Hölder, Helmhotz, Campbell, and others (for an overview, see Tal (2013)). In their characterisation, a representational measurement procedure allows one to make two formal statements,

> 'a representation theorem, which asserts the existence of a homomorphism ϕ into a particular numerical relational structure, and a uniqueness theorem, which sets forth the permissible transformations $\phi \mapsto \phi'$ that also yield homomorphisms into the same numerical structure. A measurement procedure corresponds in the construction of a ϕ in the representation theorem.' (Krantz *et al.* (1971, 12).

In the received interpretation of RTM, we thus speak of a homomorphism between an *empirical relational structure* (ERS) and a *numerical relational structure* (NRS) characterising a measurement. For example, for simple length measurement, we might want to specify the ERS as $\langle X, \circ, \succcurlyeq \rangle$, where X is a set of rods, $\circ$ is a concatenation operation, and $\succcurlyeq$ is a comparison of length of rods. If the concatenation and comparison of rods satisfy a number of conditions, there is a homomorphism into a NRS $\langle \mathbb{R}, +, \geq \rangle$, where $\mathbb{R}$ denotes the real numbers, $+$ addition operations, and $\geq$ comparison operations between real numbers. As mentioned above, the existence of such homomorphism is asserted by a *representation theorem*.

The exact characterisation of what kind of scale a given measurement procedure yields is given by *uniqueness theorems* which specify the permissible transformations of the numbers. More formally, uniqueness theorems assert that '... a transformation $\phi \mapsto \phi'$ is permissible if and only if ϕ and ϕ' are both homomorphisms ... into the *same* numerical structure ...' (Krantz *et al.*, 1971, 12). Following Stevens (1946), a distinction is usually made between nominal, ordinal, interval and ratio scales. Nominal scales allow only for one-to-one transformations. Ordinal scales allow monotonic increasing transformations of the form $\phi \mapsto f(\phi)$. Interval scales allow for affine transformations of the form $\phi \mapsto \alpha\phi +$

$\beta, \alpha > 0$. Ratio scales allow for multiplicative transformation of the form $\phi \mapsto \alpha\phi, \alpha > 0$.

Particular variants of construction such scales have emerged, such as extensive, conjoint, bisection and difference measurement (reviewed in Suppes (2002, 63ff.)). Representations in extensive measurement specify procedures that make use of the addition of magnitudes, such as in measuring physical magnitudes of mass and length. Bisection measurement gives representations by using the operation of identifying a midpoint in an interval. Conjoint measurement representations allow the combinations of magnitudes or properties, such as when measuring the intensity and frequency of a phenomenon. In difference measurement, representations capture the intensity of a particular property or relation. The three books by Krantz *et al.* (1971), Suppes *et al.* (1989) and Luce *et al.* (1990) contain a collection of mathematical results that pertain to these different measurements.

In the received interpretation, RTM takes measurement to consist in constructing homomorphisms of this kind:

> '[. . .] measurement may be regarded as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful.' (Krantz *et al.*, 1971, 9)

I call this the received interpretation of RTM because on the one hand it is close to the sparse interpretative remarks given in what is now the authoritative statement of RTM, and on the other hand it also suggests that RTM constitutes a candidate theory of measurement.

RTM as a candidate theory of measurement has been met with a fair share of criticism in the literature. Since the purpose of this paper is to sidestep rather than to engage this criticism, I will not go into detail about it. The critics focus mainly on the fact that RTM advances an abstract conception of measurement that is not connected to empirical work as closely as it should be: it reduces measurement to representation, without specifying the actual process of measuring something, and problems like measurement error and the construction of reliable measurement instruments are ignored (Michell, 1990; Decoene *et al.*, 1995; Michell, 1995; Boumans, 2007; Reiss, 2008). My proposed interpretation will sidestep these criticisms.

From this can be concluded that some critics regard RTM to be limited in serving as a theory of measurement. Where does this leave us with regards to RTM? I will not answer this question directly, as I will not assess the merits of RTM as a full-fledged theory of measurement in this paper. Rather, I argue in the following that a slightly modified interpretation of RTM can help to make it useful for a number of crucial exercises in several different fields.

3 Qualitative Relational Structures

In this section, I outline the main elements of the new interpretation of RTM. The new interpretation proceeds in two simple moves: firstly, I start from viewing RTM as a library of mathematical theorems of a certain kind. Secondly, I change the standard interpretation of the domain of the representation theorems: instead of empirical relational structures, I will interpret them more generally as *qualitative relational structures*. With these two moves completed, we can use the theorems in RTM in a greater variety of contexts. Before elaborating on the latter, I will now explain the two moves in greater detail.

Firstly, RTM is simply viewed as a library of theorems. That is to say, in the following, the term RTM will refer to the theorems in the three books that contain the authoritative statement of RTM (Krantz *et al.*, 1971; Suppes *et al.*, 1989; Luce *et al.*, 1990). Interestingly, there is relatively little by way of 'measurement' interpretation of the theorems in these three books, even though RTM is still considered to be one of the main theories of measurement, if not the dominant one (Boumans, 2008). The interpretation of the mathematical structures as referring to measurement is by and large confined to a few smaller sections in those three books (such as Chapter 1 of Krantz *et al.* (1971) and some sections of Chapter 22 in Luce *et al.* (1990)). More importantly, the idea that RTM is a full-fledged theory of measurement appears in the dozens of articles in which the different theorems have been initially been presented (extensively referenced in Krantz *et al.* (1971)). As perhaps the most poignant example of these articles, consider Davidson *et al.* (1955), in which we find extensive discussion of how the proposed theorems might make measurement in psychology and economics more 'scientific'. On the one hand, this suggests that the main proponents of RTM have undeniably intended it as a full-fledged theory of measurement. At the same time, the theorems in the three volumes cited above can also be seen separate from that. The first move of the new interpretation is to do just that and hence to consider RTM as the collection of mathematical theorems of a certain kind.

Secondly, from a mathematical point of view, the representation and uniqueness theorems in RTM simply characterise mappings between two kinds of structures, with one of these structures being associated with properties of numbers, and the other with qualitative relations. In the case of simple length measurement, the concatenation operation and the ordering relation are interpreted as actual comparisons between rods. Yet, since the theorem just concerns the conditions under which the concatenation operation and the ordering relation can be represented numerically, it is possible to furnish an even more general interpretation of what hitherto has been called ERS, the empirical relational structure. This more general interpretation is to replace the specific idea of ERS

structure with that of a QRS, a qualitative relational structure.

Reinterpreting the empirical relational structure $\langle X, \circ, \succcurlyeq \rangle$ as a *qualitative relational structure* (QRS) does not require any change, addition or reconsideration of the measurement and uniqueness theorems in RTM. Indeed, all what is needed in order to apply the latter is that there is:

- a set of well specified objects in the mathematical sense: that we have clear membership conditions for the set X. Mathematically, RTM theorems do not require that the objects have empirical content.
- well-defined qualitative relations, such as $\circ$ and $\succcurlyeq$. Mathematically, RTM theorems do not require that these relations are interpreted empirically, i.e. that we can concatenate physical objects, or compare objects empirically.

The new interpretation of RTM hence sees it only as a collection of theorems that investigate how a QRS $\langle X, \circ, \succcurlyeq \rangle$ can be mapped into a NRS $\langle \mathbb{R}, +, \geq \rangle$. It thus clearly sidesteps any of the criticisms of RTM in its received interpretation, since these criticisms were directed at RTM as a full-fledged theory of measurement, and focused on how RTM theorems apply to empirical relations.

4 Advantages of the New Interpretation

The new interpretation allows us to apply RTM theorems to any concept that we might care to investigate with regards to its potential for numerical representation. I will discuss two desiderata, firstly investigating the numerical representability of concepts, and secondly investigating possibilities of backwards engineering of foundations, before turning to defend the interpretation against two possible objections.

4.1 Numerical Representability of Concepts

With the new interpretation of RTM, we can also ask what kind of qualitative relations between imagined or idealised objects could be represented numerically. This is helpful in areas of inquiry in which there are no (or not yet developed) well-formed empirical concepts, and where there is a lack agreement on a number of basic questions.

Interpreting RTM theorems as specifying conditions of mappings between QRS and NRS, we can use them to speculate about possible numerical representations of abstract properties of abstract concepts. What is required for this are simply concepts that specify a well-defined set of objects and qualitative relations. There are some indications that RTM theorems are already used in such a way in different areas of inquiry.

Take the case of linguistic analysis of interadjective comparisons (Bale, 2008). in this field, it is investigated how we can make sense of statements such as 'x is P-er than y is Q', with van Rooij (2011) applying RTM theorems to such statements to investigate whether properties P and Q of objects x and y are numerically representable, what possible scale properties such representations can fulfil, and hence in how far interadjective comparisons can be meaningful. In short, he uses RTM theorems to investigate to what extent abstract properties that are described by adjectives can be numerically represented, and in what way they can be compared.

Another case can be found in recent philosophical investigations concerning the foundations of rational choice. Traditionally, rational choice has used RTM-style theorems in their received interpretation, investigating how preferences can be represented numerically, under the assumptions that preferences are nothing but, or closely linked to, observable choice behaviour (Davidson *et al.*, 1955). Yet, with cases of preference reversal and change, and investigations into how conflicting desires and beliefs can be captured by preferences, recent philosophical literature in rational choice theory has used RTM theorems without presupposing such close empirical links (see, for instance, Bradley (2009a), Bradley (2009b), Dietrich and List (2009), List and Dietrich (2013)). These articles investigate the determinants and changes of preferences by depicting them in an abstract way, leaving open how they are linked to observable or empirically testable entities or events.

These cases show, that some fields have already – unwittingly, or implicitly – adopted a more liberal interpretation of RTM theorems and tailored them to their needs. This suggests that the new interpretation of RTM fits well with scientific practice in some areas. At the same time, there are many more areas in which the new interpretation could help to structure similar exercises.

Consider, for instance the notorious case of personal identity over time in metaphysics (Noonan, 1989; Olson, 2002). As is well known, there is widespread disagreement over how to characterise personal identity over time, and the relevant literature is strewn with paradoxes and thought experiments that seem to pose insurmountable problems for any theory of personal identity over time. At the same time, these theories have undoubtably advanced our understanding of their subject. As a brief sketch how RTM theorems could be helping in further investigating theories of personal identity over time, consider Parfit (1984) who maintained to view persons as sets of temporal selves, and that personal identity consists in connectedness, which in turn is determined by an appropriate degree of psychological continuity between selves. To investigate to what extent the concept of a degree of psychological continuity can be represented numerically, we can interpret a QRS $\langle X, \circ, \succcurlyeq \rangle$ in which X is a set of temporal selves, and $\circ$ and $\succcurlyeq$ are operations that join and compare the psychological continuity of selves. That is, we can imagine that there is

a collection of temporal selves all of which might take differing attitudes, and who might overlap in various ways with each other. It is natural investigate these comparisons with RTM theorems: do they satisfy certain conditions such that the QRS of temporal selves and comparisons can be represented by some NRS? If so, we would be able to specify a concept of psychological continuity that is numerically representable.

Following the new interpretation of RTM is therefore both in line with recent developments in fields such as linguistics and rational choice, as well as open up applications of RTM theorems in other areas of inquiry.

4.2 Backwards Engineering of Foundations

A second advantage of the new interpretation of RTM is that it affords us greater flexibility in 'backwards engineering' of foundations. All this means is that in contexts in which we operate with numbers that lack adequate conceptual and epistemic foundations, we can investigate what kinds of qualitative relations between what kinds of objects would need to exist in order to motivate the kind of numbers that are already in use. That is, we can look at areas of inquiry that use quantities that are not derived from a measurement process and investigate whether these quantities can be seen as numerical representations of qualitative relations (and, in turn, whether such a representation can help in devising a measurement process). This holds especially for the social sciences, for which Cartwright (2008) already has made the case that RTM theorems can be very useful, even though she retained their received interpretation.

On the more general interpretation of RTM put forward here, we can jointly endorse both RTM and the view that there may be concepts relevant for a given area of inquiry that may not be directly observable. The new interpretation does allow for 'backwards engineering' of conceptual and epistemic foundations in different steps. Suppose there is some area of scientific inquiry in which numbers are currently used, yet there is ambiguity about how these numbers come about, i.e. what their conceptual foundations are and what they express, such as in happiness measurement and time discount rates in economics (and similar concepts in psychology, social science, and economics). The new interpretation of RTM allows firstly to use RTM theorems for conceptual clarification, for instance by asking in how far the concept of happiness can be represented numerically. If it is possible to conceive of the concept of happiness as numerically representable in a RTM theorem, then that is a key step in the investigation of the conceptual foundations. Secondly, we can also investigate epistemic (or evidential) foundations, by adopting the traditional interpretation of RTM, asking whether there are indeed empirical relations that sustain both the concept of happiness and a theorem in RTM.

As another example for backwards engineering of foundations, I highlight the con-

tentious issue of time discounting in economics. The idea of discounting the future – to slightly devalue the utility of future events – was introduced into economic theory by a number of different authors, but most importantly by Ramsey (1928) and Samuelson (1937, 1939). These authors provided the idea of a discount rate with which future value would be weighted, which became common practice in economics. Only much later, the idea of discounting the future was investigated in a more thorough way, providing an axiomatic basis for it, notably by Koopmans (1960). The result of these developments is that time discounting is up until this day a contentious subject in economics, with many applications requiring the use of time discount rate, but with considerable ambiguity and controversy about descriptive and normative questions about time discounting remaining (see Frederick *et al.* (2002) for an overview). Put simply, most economic theorists and practitioners live by Ramsey's (1928) dictum that time discounting is 'a practice which is ... indefensible [...] we shall, however, ... include such a rate of discount in some of our investigations.'

From the point of view of the new interpretation of RTM, this practice can be seen as ascribing numbers (discount factors) to future prospects. Naturally, the question arises whether these numbers are meaningful, i.e. whether they correspond to quantities or empirical relations. However, since future prospects are not naturally empirical entities (they are, at best, propositions), RTM in its received interpretation would be inapplicable – or only applicable in as far as one can formulate future prospects as propositions that can be subjected to observable choice behaviour. Yet, many descriptive and normative questions about time discounting go beyond that, such as those that have to do with future generations, and those for which it is impossible to sensibly formulate choice-ready propositions, such as 'branching cases' considered in theories of personal identity over time.

In this context, RTM theorems have been used by some authors in both economics and philosophy, such as Fishburn and Rubinstein (1982); Ok and Masatlioglu (2007); Heilmann (2008), to investigate how far time discounting factors can be seen as numerical representations of concepts that are important about the future, such as impatience, different types of uncertainty, and ethical judgements of various kinds. Yet, most of these efforts have been bound by the received interpretation of finding corresponding empirical structures. Investigating hypothetical scenarios and comparisons between them is something that is only possible once the new interpretation of RTM is adopted.

More generally, any case in which one is confronted with the uses of numbers or the supposition of quantitative concepts can potentially investigated with the tools of RTM – if the new interpretation is adopted.

4.3 Objections

I now turn to two objections against the new interpretation. Firstly, from the point of view of proponents of RTM as a full-fledged theory of measurement, the new interpretation might seem as 'giving in too quickly'. While it is true that the new interpretation does not endorse RTM as a full-fledged theory of measurement, it is not inconsistent with it. Rather, it spells out in what way RTM provides a useful tool, regardless of what general account of measurement they are invested in. Since numerical representability of concepts is a difficult part in many areas of measurement, rehabilitating RTM as a tool for those areas is a project that should appeal to the RTM supporter. Whether or not to additionally claim that RTM is useful beyond the two uses spelled out in the preceding sections is simply a different question that is independent from the new interpretation advanced here. Indeed, interpreting the relations as necessarily having empirical content is a special case of the more general interpretation I put forward here.

Secondly, it is possible to question whether the interpretation advanced here does make a big difference to RTM and its use. I think such an objection underestimates both the problems commonly associated with RTM in its received interpretation and their possible consequences. Since RTM is widely and heavily criticised as a theory of measurement (see Section 2), there is a real danger that it will be dispensed altogether. The consequence of that would, in turn, be that exercises as described in Section 4.1 and 4.2, already carried out in some fields, would be without an account that underpins them. Moreover, the additional perspectives to discuss the numerical representability of properties of hypothetical entities (such as temporal selves in metaphysics) or reasons to discount the future which cannot be easily grounded in empirical relations hang on adopting the new interpretation.

5 Conclusions

In this paper, I have proposed to interpret the Representational Theory of Measurement in a new way, namely as a library of theorems that investigate the numerical representability of qualitative relations. Such theorems are useful tools for concept formation which, in turn, can be seen as one crucial aspect of measurement for a broad range of cases in linguistics, rational choice, metaphysics, and the social sciences. I have suggested that it is already part of scientific practice to use RTM theorems in such a way, and have suggested that there are more cases to which they could be fruitfully be applied.

Acknowledgements

Many thanks in particular to Eran Tal for many helpful comments, as well as to Constanze Binder, Marcel Boumans, Aki Lehtinen, Luca Mari, F.A. Muller, Julian Reiss, Jan-Willem Romeijn, and participants at the 2012 Arctic Workshop on Measurement in Rovaniemi and the 2013 OZSW Conference of the Dutch Research School of Philosophy in Rotterdam. Work on this article has been supported by a Marie Curie Career Integration Grant #PCIG10-GA-2011-303900 from the European Union and a VENI grant #275-20-044 from the Netherlands Organisation for Scientific Research (NWO).

References

Adams, E. W. (1966). On the nature and purpose of measurement. *Synthese*, **16**, 125–169.

Bale, A. (2008). A universal scale of comparison. *Linguistics and Philosophy*, **31**, 1–55.

Boumans, M. (2007). *Measurement in Economics: A Handbook*. AP Elsevier.

Boumans, M. (2008). Measurement. In S. N. Durlauf and L. E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke.

Bradley, R. (2009a). Becker's thesis and three models of preference change. *Politics, Philosophy and Economics*, **8**(2), 223–242.

Bradley, R. (2009b). Preference kinematics. In T. Grüne-Yanoff and S. O. Hansson, editors, *Preference Change*, pages 221–242. Theory and Decision Library A, Springer.

Cartwright, N. (2008). In praise of the representation theorem. In M. Frauchiger and W. K. Essler, editors, *Representation, Evidence, and Justification: Themes from Suppes*, pages 83–90. Ontos Verlag.

Cartwright, N. D. and Chang, H. (2008). Measurement. In S. Psillos and M. Curd, editors, *The Routledge Companion to Philosophy of Science*, pages 367–375. New York: Routledge.

Davidson, D., McKinsey, J. C. C., and Suppes, P. (1955). Outlines of a formal theory of value, I. *Philosophy of Science*, **22**(2), pp. 140–160.

Decoene, S., Onghena, P., and Janssen, R. (1995). Representationalism under attack. Review of An introduction to the logic of psychological measurement, by J. Michell and Philosophical and foundational issues in measurement theory, by C. Wade Savage and P. Ehrlich. *Journal of Mathematical Psychology*, **39**(2), 234 – 242.

Dietrich, F. and List, C. (2009). A model of non-informational preference change. *LSE Choice Group Working Paper Series*, **5**(1).

Fishburn, P. C. and Rubinstein, A. (1982). Time preference. *International Economic Review*, **23**(3), 677–94.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, **40**(2), 351–401.

Heilmann, C. (2008). Measurement-theoretic foundations of time discounting. *LSE Choice Group Working Paper Series*, **4**.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica*, **28**(2), 287–309.

Krantz, D. H., Luce, R. D., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Mineola: Dover Publications.

List, C. and Dietrich, F. (2013). Where do preferences come from? *International Journal of Game Theory*, **42**(3), 613–637.

Luce, R. D., Krantz, D. H., Tversky, A., and Suppes, P. (1990). *Foundations of Measurement Volume III: Representation, Axiomatization, and Invariance*. Mineola: Dover Publications.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Erlbaum.

Michell, J. (1995). Further thoughts on realism, representationalism, and the Foundations of measurement theory. Author's Response to Review by Decoene et al. of An introduction to the logic of psychological measurement. *Journal of Mathematical Psychology*, **39**(2), 243–247.

Noonan, H. W. (1989). *Personal Identity*. Routledge.

Ok, E. A. and Masatlioglu, Y. (2007). A theory of (relative) discounting. *Journal of Economic Theory*, **137**, 214–45.

Olson (2002). *Personal Identity*. The Stanford Encyclopedia of Philosophy (Fall 2002 Edition), E. Zalta.

Parfit, D. (1984). *Reasons and Persons*. Clarendon.

Ramsey, F. P. (1928). A mathematical theory of saving. *Economic Journal*, **38**(152), 543–59.

Reiss, J. (2008). *Error in Economics: Towards a More Evidence-based Methodology*. Routledge.

Russell, B. (1903). *The Principles of mathematics*. New York: Cambridge University Press.

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, **4**, 155–61.

Samuelson, P. (1939). The rate of interest under ideal conditions. *The Quarterly Journal of Economics*, **53**(2), 286–97.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**(2684), 677–680.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. Stanford: CSLI Publications.

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1989). *Foundations of Measurement Volume II: Geometrical, Threshold, and Probabilistic Representations*. Mineola: Dover Publications.

Swistak, P. (1990). Paradigms of measurement. *Theory and Decision*, **29**(1), 1–17.

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass*, **8**(12), 1159–1173.

van Rooij, R. (2011). Measurement and interadjective comparison. *Journal of Semantics*, **28**, 335–358.

Theory Testing and Implication in Clinical Trials

March 1, 2014

Abstract

John Worrall (2010) and Nancy Cartwright (2011) argue that randomized controlled trials (RCTs) are "testing the wrong theory." RCTs are designed to test inferences about the causal relationships in the study population, but this does not guarantee a justified inference about the causal relationships in the more diverse population in clinical practice. In this essay, I argue that the epistemology of theory testing in trials is more complicated than either Worrall's or Cartwright's accounts suggest. I illustrate this more complex theoretical structure with case-studies in medical theory testing from (1) Alzheimer's research and (2) anti-cancer drugs in personalized medicine.

1 Introduction

John Worrall (2010) and Nancy Cartwright (2011) have both argued that there is a mismatch between the theory being tested in a randomized controlled clinical trial (RCT) and the theory that medical practitioners are actually interested in. Worrall describes this as the problem of external validity: An RCT may support internally valid inferences about the causal relationships in the study population, but this does not guarantee a justified inference about the causal relationships in the target population of interest—i.e., the usually more diverse patient population that physicians actually encounter in the clinic. Since it is the causal relationship between treatment and patient outcome in this more diverse population that we ultimately care about, the RCT is "testing the wrong theory" (p.361). Or as Cartwright puts it: "an RCT supports only an 'it-works-somewhere' claim," but what we need are justified "it will work for us" claims (p.1401).

There is something fundamentally correct in both Cartwright's and Worrall's arguments: Many of the experimental components used in RCTs are designed to secure internal validity at the expense of external validity. Yet, the epistemological relationship between translational clinical trials (whether randomized or not) and the underlying scientific theories is more complicated and, in some ways, more subtle than either of their accounts would suggest. In this essay, I will illustrate this more complex theoretical structure through two examples, drawn from trials in Alzheimer's research and anti-cancer drugs for personalized medicine. I argue that the more complicated epistemology vis-a-vis theory testing revealed in these

cases illuminates how Worrall's and Cartwright's philosophical conclusion relies on an overly narrow conception of what trials can show.

2 Testing the Wrong Theory

Let us begin, as Worrall does, by imagining an RCT that is evaluating some new intervention S as a treatment for some condition C in a sample population P. Supposing that the trial is positive, what can we now conclude? Worrall cautions that we should not conclude the "dangerously vague" claim that "S is effective for treating C". Rather, the trial's result only warrants a narrower claim that "S when administered in a very particular way to a very particular set of patients for a particular length of time is more effective than some comparator treatment" (Worrall, 2010, p.361).

In other words, if we assume that the trial is internally valid, then it only justifies the claim that "S was effective for condition C in population P". But the practicing physician needs to know how or whether this effectiveness claim generalizes to the variations on these parameters that she is likely to encounter in the clinic. Does the relationship between S and C also hold for more elderly patients (who are typically excluded from trials)? Does it hold if we modify the dose and schedule to accommodate patients with co-morbidities or concomitant medications? Is S also an effective intervention for treating the related condition, C^*? All of these questions speak to the problem of external validity: An RCT may demonstrate a causal relationship between S and C in the studied population P. But what we want to know is whether this relationship holds for $\mathbb{S}$ and $\mathbb{C}$ in the population $\mathbb{P}$—where $\mathbb{S}$, $\mathbb{C}$, and $\mathbb{P}$ are the respective sets of plausible variations on the intervention, condition, and patient populations that the physician encounters in clinical practice.

And as Worrall goes on to point out, there are often reasons to doubt that P is a good representative (or perhaps even a member) of the relevant clinical set $\mathbb{P}$. For example, the limited observation time in an RCT weakens inferences about the safety of treatments for chronic diseases, such as diabetes or arthritis. Even though large RCTs for these conditions will usually include a few years of follow-up, the general patient population is likely to be taking these medications for decades, and the RCT has not provided any evidence about such long-term effects. Similarly, some RCTs will include time-consuming procedures or expensive diagnostics as a part of the treatment regimen. Insofar as these same procedures or tests are unavailable to the clinician (due to excessive cost, timing, or feasibility), then the results of the RCT can fail to provide clinically relevant evidence about how the intervention should be used in practice.

Cartwright (2011) extends this line of argument with some additional analytic resources. She draws a distinction between experimental "vouchers" and "clinchers" (cf. Cartwright, 2007). A "voucher" is an experiment that renders its conclusion more probable, whereas a "clincher" is an experiment that deductively implies its conclusion. As she defines it, an ideal RCT "where all the requisite premises are met" is a clincher—and this is presumably why the RCT sits atop the hierarchy of evidence-based medicine.

But what are these "requisite premises" on which the "clinching" depends? Cartwright

enumerates three of them: (1) probabilistic dependence calls for causal explanation; (2) all causal features in the study population P relevant to the outcome, except for the treatment, are equally distributed between the treatment and control arms; and (3) the experimental treatment S is the only plausible explanation for the observed difference in outcome between the arms (p.1400).[1]

Let us set premise (1) aside here, since discussing the necessity (or not) of causal explanations for probabilistic dependence will take us too far afield. Premises (2) and (3) deserve some attention, however. As Cartwright acknowledges, RCTs are explicitly designed to satisfy these two claims. Random treatment allocation, in particular, is typically justified for exactly this reason: It controls for all known and unknown confounders in the study population. Restrictive eligibility criteria, strict treatment protocols, allocation concealment, and blinded outcome assessment are also characteristic features of the RCT—all of which are introduced to eliminate the influence of bias, and thereby increase our confidence in premises 2 and 3.

But as Cartwright observes, these methodological steps also render the RCT less like clinical practice. In the clinic, physicians will often modify a treatment's dose or schedule, or even switch patients from one drug to another, in the face of observed inefficacy, adverse reactions, or side-effects. Patients are also neither blinded to their prescribed treatment nor prescribed a treatment randomly. And just as Worrall argued, the clinical patient population $\mathbb{P}$ is usually far more diverse than the study population P. Each of these differences between the RCT and practice weaken the inference (i.e., generalizability) from causal claims about what occurred in the study to causal claims about what will occur in the clinic.

To resolve this problem, Cartwright argues that we need justified claims about the causal "capacities" of our treatment S—that is, *theoretical* warrant for thinking that S is a good representative of $\mathbb{S}$, C a good representative of $\mathbb{C}$, and P a good representative of $\mathbb{P}$. As she puts it, we need the theoretical understanding of "why the treatment should have the power to produce the outcome". Unfortunately, all we get from an RCT is evidence that the S can "work somewhere," but this is not the same as having a justified theory of why we should expect that "it will work for us" (p.1401).

3 External Validity and Underdetermination

As I suggested above, there is something fundamentally correct in Worrall's and Cartwright's arguments. The RCT is typically designed to ensure internal validity—to "clinch" (to use Cartwright's term) the causal hypothesis that the experimental treatment S is efficacious against condition C in the study's population P. But as we just saw, the steps taken to guarantee the validity of this inference will often weaken its external validity. And this trade-off

[1] Cartwright does not use these variables in her formulation. I introduce them here to better accord with the Worrall discussion above, but I trust it does no violence to her account. I have, however, weakened her third premises. Her original wording—"the *only explanation possible* is that the treatment caused the outcome" (p.1400, emphasis added)—is far too strong and inconsistent with any credible account of RCT methodology, see for example Shadish et al. (2002) or Friedman et al. (2010).

on internal versus external validity leads us to Worrall and Cartwright conclusions: RCTs are not testing the right theory. They are not telling us what we need to know.

This conclusion is consonant with others in the medical literature who have called for more "comparative effectiveness" trials. For example, Tunis et al. (2003) argue that too many RCTs evaluate the new drug against a placebo comparator, even when there is already a proven effective treatment available. But if the only evidence about some new drug, A, is its superiority to placebo, this does not provide clinicians with sufficient knowledge about whether they should be prescribing A over the old standard of care. It can also be traced back to Schwartz and Lellouch (1967), who drew a distinction between *pragmatic* and *explantory* trials. A pragmatic trial is conducted under conditions similar to clinical practice and seeks to answer a question about medical decision-making, e.g., "Which treatment should physicians use in practice?" Whereas an explanatory trial is conducted under "ideal" scientific conditions and seeks to answer a question about scientific understanding, e.g., "What is the true biological effect of drug S?" Their philosophical point is similar to Worrall's and Cartwright's: Different experiments can address different theories, so we should be conducting experiments that answer the questions we are actually interested in. If we want to answer clinically relevant questions, then we should be conducting pragmatic trials that maximize external validity.

I take it, however, that Worrall and Cartwright are not simply echoing Schwartz and Lellouch. They seem to be saying something stronger—namely, that RCTs are testing the "wrong theory." But what should we make of this claim? It is certainly true that most RCTs adopt some version of Neyman-Pearson hypothesis testing, and are therefore, strictly speaking, tests of a single hypothesis (or single theory, if you prefer). Yet, it would be a mistake to think that an RCT has no further theoretical importance. This much follows trivially from underdetermination: Multiple scientific theories are involved in the design of an experiment and therefore multiple scientific theories are implicated by the evidence produced—e.g., theories about the therapeutic class (of which the drug is just one member) and its relationship to disease modification; theories about the diagnostic assays and their relationship to the disease prognosis; theories of disease ontology and pathophysiology. A negative RCT, for example, does not necessitate that the researchers reject any causal link between the treatment S and the condition C. Perhaps S will be effective against C in a slightly different population P^*. The essential point of underdetermination is that there are always auxiliary hypotheses or other theoretical modifications that can be made to accommodate the evidence.[2]

Since it seems unlikely that Worrall and Cartwright would object to the relevance of theoretical underdetermination in RCTs, perhaps we ought to interpret their conclusion differently. Maybe what they are really arguing is that the inferences to these other theories are not well-justified by the evidence produced in an RCT. Cartwright, in particular, has the conceptual resources to still conclude that RCTs are not "clinchers". At best, they are only "vouchers" for most of the relevant theoretical claims.

[2]Worrall discusses Duhem's problem earlier in the same article (Worrall, 2010, p.358), but overlooks its relevance for his argument on theory testing. See also Anderson (2006), Howick (2009), and Hey and Weijer (2013) for more details discussions of Duhem's problem and its importance for understanding the methodology of clinical trials.

But if this is really their conclusion, much of the philosophical force of their argument is lost. If the argument simply is the same as Schwarz and Lellouch's or Tunis et al.'s—that we need pragmatic trials to answer (or "vouch for") clinically relevant questions—then I agree entirely. But this does not seem consistent with much of what Worrall and Cartwright argue. Even the most pragmatic trial is still just a "voucher" for clinical effectiveness. It is asking the more clinically relevant question. Yet, if it differs in any way from the clinical setting, then on Worrall's and Cartwright's view, it does not provide the right kind of evidential support.

Indeed, Cartwright emphasizes that "RCTs do not, without a series of strong assumptions, warrant predictions about what happens in practice" (p.1400). And Worrall concludes that we should really be doing observational studies rather than RCTs (p.362). But these conclusions seem untenable. Even setting aside the obvious objection to Worrall that observational studies have their own biases and methodological limitations (and are just as subject to underdetermination), it is much to strong to demand of an experiment that it provide direct evidence or causal certainty before we can draw externally valid inferences from it. "Clinchers" may be a worthy philosophical ideal, but it does not follow that this is a plausible experimental benchmark. Demanding deductive causal certainty is to ask more of an experiment than it can plausibly provide.

4 Trials and Theoretical Implication

So where does this leave us? I agree with Worrall and Cartwright that the external validity of RCTs is limited (as it is for every experiment). Yet, the story of clinical trials and theory testing is more complicated than either of their accounts would seem to suggest. In this section, I will discuss two examples of theory testing in medical research, each of which illuminates a number of different ways in which trials have a theoretical import beyond their specific testing hypothesis.

4.1 The Amyloid Cascade

Much of Alzheimer's research has been driven by a mechanistic theory of the amyloid cascade, which posits that the characteristic neurodegeneration of Alzheimer's disease is caused by amyloid-β plaque accumulation in the brain. However, as Karran et al. (2011) describe, even as various "amyloid-centric" approaches have failed (e.g., the drugs tramiprosate, tarenflurbil, semagacestat all failed in development), the fundamental amyloid cascade theory has not been rejected. It has only been modified. They now distinguish between three different theoretical amyloid-centric strategies: reducing amyloid-β production, facilitating amyloid-β clearance, and preventing amyloid-β aggregation (p.700).

The drug company, Genentech's, anti-amyloid monoclonal antibody, crenezumab, is one such amyloid-centric drug currently undergoing clinical trials. In fact, it is being tested in two different Alzheimer's trials: One is a long-term single arm trial in patients with mild to moderate Alzheimer's symptoms; the other is a double-blind RCT testing crenezumab as a

neuroprotective agent in a genetically homogeneous population in Columbia.[3] So what are the theoretical implications of these trials?

For the single arm study, a negative result would provide evidence that crenezumab is not an effective strategy for treating Alzheimer's symptoms. It would also provide evidence that similar monoclonal antibodies are unlikely to be effective, as well as further evidence for the growing suspicion that once amyloid-β deposition has begun, removing amyloid-β is unlikely to offer any therapeutic benefit (Golde et al., 2011). Whereas a positive result would confirm both (a) that crenezumab and similar monoclonal antibodies may be viable strategies, and (b) that an amyloid-β clearance strategy is effective.

Similarly, for the RCT in Columbia, a negative result would be evidence against crenezumab's effectiveness. It would also provide disconfirming evidence that preventing amyloid-β aggregation offers any neurodegenerative protection. A positive result would confirm both of those theories: preventing amyloid-β aggregation is a viable strategy and crenezumab, in particular, is likely to be an effective treatment.

I am happy to grant a possible Cartwright objection here that neither of these trials "clinch" any of these theoretical claims. But surely the more relevant question is whether or not these trials provide sufficient evidence for informing clinical decision-making. And on that point it is instructive to observe that part of the inclusion criteria for the Columbian RCT is that all patient-subjects must be carriers of a specific gene mutation (PSEN1 E280A), which is known to cause early-onset Alzheimer's disease (cf. Belluck, 2012). Supposing that this trial has a positive result, what are clinicians justified in concluding about other patient populations at risk for Alzheimer's? Worrall's and Cartwright's arguments imply that clinicians would still lack sufficient evidence for prescribing crenezumab outside of that specific genetic population. But the theoretical warrant from these trials is not so weak. A success for crenezumab lends evidential support for a range of theoretical propositions, some of which would be sufficient to justify a clinician's decision to prescribe an approved anti-amyloid agent for her Alzheimer's patients.

And it brings us to the heart of the issue: The directly tested theory in the Columbian RCT could be thought of as resembling the narrow proposition, much as Worrall originally construed it: "Crenezumab (S) is effective for preventing the development of Alzheimer's disease (C) in the Columbian patient population possessing the PSEN1 E280A genetic mutation (P)." But this does not exhaust the theoretical relevance of the trial. Whatever its final result—but particularly if it is positive—researchers and clinicians will be in a better position to draw valid (albeit inductive) inferences about future preventative strategies against Alzheimer's. Specifically, they would be justified in inferring potential efficacy for other preventative anti-amyloid interventions ($\mathbb{S}$); extrapolating the strategy for related conditions, such as sporadic Alzheimer's ($\mathbb{C}$); or prescribing anti-amyloid medications for other patient populations at high-risk for developing amyloid-related neurodegenerative diseases ($\mathbb{P}$). To be sure, these would all still be inferences with some degree of causal uncertainty, but it does not follow from the lack of certainty that the inferences are unwarranted or unjustified. On the contrary, if an anti-amyloid strategy is shown to be effective in an RCT, it would arguably

[3] See http://clinicaltrials.gov/ct2/results?term=Crenezumab&Search=Search, retrieved February 27, 2014.

violate the physicians' duty of care to withhold the treatment.

4.2 Personalized Cancer Medicine

In many ways, the amyloid cascade and Alzheimer's case is an exemplar for the traditional model of clinical translation, where the driving theories concern the experimental drug's effectiveness and the mechanism of disease. As we saw, a new Alzheimer's drug that is successfully vetted in trials is taken to confirm the particular drug's effectiveness, the effectiveness of the strategic class, and the underlying theories of disease pathophysiology. Whereas the drug's failure can be attributed to either a problem with one of these theories, a faulty auxiliary hypothesis, or an operational error in one or more of the experiments.

The development of new personalized medicines (PM), however, is not well-characterized by this model. The goal in PM is to equip the health-care system with an array of clinically validated diagnostics, each of which would allow physicians to test their patients for the presence or absence of a particular biomarker (e.g., a genetic mutation in their tumor specimen), and then use these results to tailor decision-making about the appropriate course of treatment. If successfully implemented, these biomarker diagnostics would potentially save the health-care system billions of dollars and prevent needless patient suffering due to futile interventions.

On its face, the epistemology of PM, in some ways, better accommodates Worrall's and Cartwright's views. That is, PMs are designed to be effective in a very narrowly defined patient population—i.e., only those patients with the specific biomarker. Thus, the study population in RCTs for PM is far more likely to resemble the target population in clinical practice. However, in contrast to the traditional model of medical research and drug development, which hinges on effective therapeutic agents, the promise of PM largely depends upon the development of high-quality biomarker diagnostics. And this further complicates the theoretical implications of PM trials.

Consider the case of the alkylating agent, temozolomide. This drug was derived from the older, widely-used (although quite toxic) cancer agent, dacarbazine, and works by attaching an alkyl group to the cancer cell DNA, disrupting its growth and leading to cell death. Interestingly, despite sharing the same mechanism of action, temozolomide and dacarbazine are used in differenct cancers. Dacarbazine is approved for use against Hodgkin lymphoma and melanoma; temozolomide is approved for use against anaplastic astrocytoma and glioblastoma multiforme.

Let us label this broadly defined mechanistic theory of using alkylating cancer drugs, T_1:

T_1 Alkylating agents ($S_1 \ldots S_n$) are a viable treatment strategy for some patient populations ($P_1 \ldots P_n$) with some cancers ($C_1 \ldots C_n$).

Thus, dacarbazine and temozolomide are two of the agents in the set $S_1 \ldots S_n$, and the various cancers for which they have been approved are the members of the set $C_1 \ldots C_n$. One of the challenges in cancer treatment is that the patient population that benefits from a particular agent is not fully determined by their cancer-type. For example, not all patients with glioblastoma will benefit from temozolmide therapy. And this is where diagnostic biomarker assays

come into play. Indeed, the theory underlying all of PM is that there are genetic markers in a patient's tumor which can predict whether or not they are likely to benefit from a treatment.

One of the proposed biomarkers for temozolomide is the DNA repair gene O-6-methylguanine-DNA methyltransferase, typically abbreviated as "MGMT". A landmark study by Hegi et al. (2004) identified a positive correlation between patient tumor response to temozolomide therapy and high levels of methylated MGMT expression in their tumor specimens. Their conclusion can be characterized by the more specific theoretical hypothesis, T_2:

T_2 Temozolomide chemotherapy (S_g) is most likely to be effective against glioblastoma tumors (C_g) for those patients whose tumor specimens express high levels of methylated MGMT (P_g).

We can think of T_2 as a sub-theory of T_1, since it describes a relationship among a single triad of the treatment-condition-population parameters. And although T_1 is uncontroversial and has already been taken up in clinical practice, T_2 is still being rigorously evaluated in trials.[4] But just as we saw with the amyloid cascade theory and crenezumab, a positive or negative result in any of these trials has theoretical implications for both T_1 and T_2.

Yet, many of these trials have an additional dimension of uncertainty derived from the predictivity of the diagnostic assay (or assays) used to assess the methylated MGMT biomarker. There are multiple techniques that can be used to determine the level of methylated MGMT in a specimen and these different techniques do not all discriminate the glioblastoma patients in the same way. In effect, they each define the target population P_g differently. One recent study, for example, compared the sensitivity and specificity of a methylation-specific polymerase chain reaction (MS-PCR) assay, which amplifies the relevant CpG islands of the tumor specimen's DNA, against an immunohistochemistry staining (IHC) assay, which assesses the reactivity of tumor cells against a specific antibody (Lechapt-Zalcman et al., 2012). They found that although both assays positively predicted benefit from temozolomide therapy, the agreement between them was only about 70%. That is, 30% of the samples tested positive for methylated MGMT on one test, but negative on the other.

This makes the recommendation for clinical practice more problematic. What is the true patient population for our theory T_2? Is it the patients whose samples test positive on MS-PCR or IHC? A clinician's decision to recommend temozolomide now hinges, in part, on their selection of assay.

Lechapt-Zalcman et al. (2012) attribute this discrepancy largely to false-positives with the IHC, which on its face, would seem to suggest that MS-PCR is the better assay for defining the population P_g (p.4553). But they also note that the accuracy of MS-PCR depends upon high-quality cryopreserved tumor specimens, which is expensive and not widely available in the clinical setting (p.4552). Thus, despite its being the less accurate of the two assays, an IHC assay may be the more clinically useful diagnostic. And this brings us back to the problem of external validity. If MS-PCR is too expensive and unlikely to be used in the clinic, then the Worrall or Cartwright arguments would suggest that future trials ought to only

[4] At the time of this writing, 10 trials are registered on clinicaltrials.gov examining the implications of temozolomide and MGMT for the treatment of glioblastoma.

investigate IHC. Since IHC is the technique available to clinicians, then presumably, what clinicians want is evidence about its capacity to delineate the responding patient population P_g.

Unfortunately, even this seemingly reasonable suggestion still relies on an oversimplification of the theoretical implications in these studies. To wit, we should observe that the effective use of a diagnostic test depends on knowing its misclassification rate, i.e., the false-positive and false-negative error rates. If a gold-standard diagnostic exists—that is, a diagnostic with perfect sensitivity and specificity—then these error rates are easy to determine. One can simply compare the classification of the imperfect diagnostic, e.g., IHC, to the classification according to the gold-standard. Of course, in practice, there are no gold-standards. Every diagnostic is imperfect. However, there are validated techniques for accurately estimating the error rates of a test on the basis of multiple diagnostics. In essence, these are robustness strategies, which use multiple independent (or sometimes conditionally dependent) tests in order to arrive at estimates for the error rates of each individual diagnostic (Joseph et al., 1995).

And indeed, relying on multiple diagnostics is precisely the strategy adopted in some of the more recent temozolomide studies (cf. Lalezari et al., 2013). Given that IHC is known to be inaccurate, researchers in these trials can use other, more accurate diagnostics (e.g., MS-PCR, pyrosequencing) in combination with IHC in order to derive better estimates for the error rates when using a single IHC diagnostic test. These estimates can then be used by clinicians, who may only have access to one diagnostic method, to make informed decisions about their patient's true biomarker status and potential benefit from temozolomide therapy.

The essential philosophical point here is that these rigorous biomarker studies do have weaker external validity. They employ multiple diagnostics and robustness strategies, which may be unavailable or unwieldy in clinical practice. Yet, their use of multiple diagnostics toward a more robust theoretical understanding of the various individual techniques is precisely what makes them informative for clinical practice. Contrary to Cartwright's claim, these explanatory (or "ideal") trials are addressing the clinically relevant theoretical question—"What is the accuracy of IHC for predicting response to temozolomide?" This is exactly the kind of information that clinicians need to know in order to make the most of PM in cancer.

5 Conclusion

What theory or theories are tested in clinical trials? I have argued here that the answer to this question is more complicated than suggested by either Worrall (2010) or Cartwright (2011) in their critiques of RCTs. Their emphasis on the problem of external validity is helpful, insofar as it draws greater attention to the need for studies that address clinically relevant questions. But their stronger conclusion against the theoretical warrant provided by RCTs relies on a significant oversimplification of trial epistemology.

As the problem of underdetermination entails, there are many theoretical implications of trials. The focal testing hypothesis of the form "Treatment S is effective for condition C in population P" is but one of the many theoretical claims that can be justifiably confimred,

modified, or refuted in light of an trial's result. RCTs also generate evidence that is relevant for general theories about the viability of the mechanistic strategy, or the underlying pathophysiological theories of the disease, or the theories concerning biomarkers, diagnostic assays, and the predictive relationship that these bear to patient prognosis and treatment. All of these moving theoretical parts are potentially implicated. To suggest otherwise assumes an overly narrow and untenable view about what RCTs can show.

References

Anderson, J. A. (2006). The ethics and science of placebo-controlled trials: Assay sensitivity and the duhem-quine thesis. *Journal of Medicine and Philosophy 31*, 65–81.

Belluck, P. (2012). New drug trial seeks to stop alzheimers before it starts. *The New York Times*. May 15.

Cartwright, N. (2007). Are rcts the gold standard? *BioSocieties 2*(1), 11–20.

Cartwright, N. (2011). The art of medicine: A philosopher's view of the long road from rcts to effectiveness. *The Lancet 377*, 1400–1401.

Friedman, L. M., C. Furberg, and D. L. DeMets (2010). *Fundamentals of clinical trials* (4 ed.). Springer.

Golde, T. E., L. S. Schneider, and E. H. Koo (2011). Anti-aβ therapeutics in alzheimer's disease: the need for a paradigm shift. *Neuron 69*(2), 203–213.

Hegi, M. E., A.-C. Diserens, S. Godard, P.-Y. Dietrich, L. Regli, S. Ostermann, P. Otten, G. Van Melle, N. de Tribolet, and R. Stupp (2004). Clinical trial substantiates the predictive value of o-6-methylguanine-dna methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clinical Cancer Research 10*(6), 1871–1874.

Hey, S. P. and C. Weijer (2013). Assay sensitivity and the epistemic contexts of clinical trials. *Perspectives in Biology and Medicine 56*(1), 1–17.

Howick, J. (2009). Questioning the methodologic superiority of 'placebo' over 'active' controlled trials. *The American Journal of Bioethics 9*, 34–48.

Joseph, L., T. W. Gyorkos, and L. Coupal (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology 141*(3), 263–272.

Lalezari, S., A. P. Chou, A. Tran, O. E. Solis, N. Khanlou, W. Chen, S. Li, J. A. Carrillo, R. Chowdhury, J. Selfridge, et al. (2013). Combined analysis of o6-methylguanine-dna methyltransferase protein expression and promoter methylation provides optimized prognostication of glioblastoma outcome. *Neuro-oncology 15*(3), 370–381.

Lechapt-Zalcman, E., G. Levallet, A. E. Dugué, A. Vital, M.-D. Diebold, P. Menei, P. Colin, P. Peruzzy, E. Emery, M. Bernaudin, et al. (2012). O6-methylguanine-dna methyltransferase (mgmt) promoter methylation and low mgmt-encoded protein expression as prognostic markers in glioblastoma patients treated with biodegradable carmustine wafer implants after initial surgery followed by radiotherapy with concomitant and adjuvant temozolomide. *Cancer 118*(18), 4545–4554.

Schwartz, D. and J. Lellouch (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases 20*(8), 637–648.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Tunis, S. R., D. B. Stryer, and C. M. Clancy (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association 290*(12), 1624–1632.

Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice 16*, 356–362.

Bell's local causality for philosophers

Gábor Hofer-Szabó[*]
Péter Vecsernyés[†]

Abstract

This paper is the philosopher-friendly version of our more technical work (Hofer-Szabó and Vecsernyés, 2014). It aims to give a clear-cut definition of Bell's notion of local causality. Having provided a framework, called local physical theory, which integrates probabilistic and spatiotemporal concepts, we formulate the notion of local causality and relate it to other locality and causality concepts. Then we compare Bell's local causality with Reichenbach's Common Cause Principle and relate both to the Bell inequalities. We find a nice parallelism: both local causality and the Common Cause Principle are more general notions than captured by the Bell inequalities. Namely, Bell inequalities cannot be derived neither from local causality nor from a common cause unless the local physical theory is classical or the common cause is commuting, respectively.

Key words: local causality, Bell inequality, common cause

1 Introduction

Local causality is the principle that causal processes cannot propagate faster than the speed of light. This does not mean that in a physical theory subject to this principle no correlation between spatially separated events can exist; a correlation can well be brought about by a common cause in the past of the events in question. However, since all causal processes propagate within the lightcone, fixing the past of an event in a detailed enough manner, the state of this event will be fixed once and for all, and no other spatially separated event can contribute to it any more.

In a nutshell, this is the idea which becomes primary focus in John Bell's (2004) seminal papers initiating a whole research program in the foundations of quantum theory. In these papers Bell translated the intuitive idea of local causality into a probabilistic language opening the door to treat the principle in a theoretical setting and to test its experimental validity *via* the Bell inequalities derived from the principle. The logical scheme of this translation was the following: if physical events are localized in the spacetime in a certain independent way, then these events are to satisfy certain probabilistic independencies. This manual was highly intuitive, however, to apply it in a formally correct way one had to wait until the advent of a mathematically well-defined and physically well-motivated formalism which is able to integrate spatiotemporal and probabilistic concepts. Without such a framework one could not account for the (otherwise intuitive) inference from relations between spacetime regions to probabilistic independencies between, say, random variables. The most elaborate formalism offering such a general framework is quantum field theory, or its algebraic-axiomatic form, algebraic quantum field theory (AQFT).

Thus, it comes as no surprise that AQFT has soon become an important medium to pursue research on the Bell inequalities (Summers, 1987a,b; Summers and Werner, 1988; Halvorson 2007); relativistic causality (Butterfield 1995, 2007; Earman and Valente, 2014); or the closely related (see below) Common Cause Principle (Rédei 1997; Rédei and Summers 2002; Hofer-Szabó and Vecsernyés 2012a, 2013a). In

[*]Research Center for the Humanities, Budapest, email: szabo.gabor@btk.mta.hu
[†]Wigner Research Centre for Physics, Budapest, email: vecsernyes.peter@wigner.mta.hu

this paper we follow the route pioneered by the algebraists, but we do not go as far as AQFT. Our aim is simply to establish a *minimal framework* which is needed to formulate Bell's notion of local causality in a strict fashion. Thus we will borrow only a part of AQFT to represent something which we will call a *local physical theory*. A local physical theory is a formal structure integrating the two most important components of a general physical theory: spacetime structure and algebraic-probabilistic structure. Our secondary aim in this paper is to clarify the relation of Bell's local causality to such other important notions as local primitive causality, Common Cause Principle and the Bell inequalities.

There is a renewed interest in a deeper conceptual and formal understanding of Bell's notion of local causality. Travis Norsen illuminating paper on local causality (Norsen, 2011) or its relation to Jarrett's completeness criterion (Norsen, 2009); the paper of Seevinck and Uffink (2011) aiming at providing a 'sharp and clean' formulation of local causality; or Henson's (2013) paper on the relation between separability and the Bell inequalities are all examples of this inquiry. Our research runs parallelly in some respect to these investigations and we will comment on the points of contact underway.

In Section 2 we fix our mathematical framework, called local physical theory and list some important relativistic causality principles. In Section 3 we formulate Bell's notion of local causality in a local physical theory. In Section 4 we compare local causality with the Common Cause Principle and relate both to the Bell inequalities. We conclude the paper in Section 5.

This paper is the philosopher-friendly version of our more detailed and more technical work (Hofer-Szabó and Vecsernyés, 2014). Many points (such as local causality in a non-atomic local physical theory; local causality in stochastic dynamics; its complex relation to other locality and causality concepts, etc.) which are treated in a more conceptual way here obtain a more detailed mathematical analysis there. We will not refer to these results point-by-point in the paper.

2 What is a local physical theory?

First we set the framework, called local physical theory, within which probabilistic and spatiotemporal notions can be treated in an integrated way.

Definition 1. A $\mathcal{P}_{\mathcal{K}}$-*covariant local physical theory* is a net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ associating algebras of events to spacetime regions which satisfies *isotony*, *microcausality* and *covariance* defined as follows (Haag, 1992):

1. *Isotony.* Let $\mathcal{M}$ be a globally hyperbolic spacetime and let $\mathcal{K}$ be a covering collection of bounded, globally hyperbolic subspacetime regions of $\mathcal{M}$ such that $(\mathcal{K}, \subseteq)$ is a directed poset under inclusion $\subseteq$. The net of local observables is given by the isotone map $\mathcal{K} \ni V \mapsto \mathcal{A}(V)$ to unital C^*-algebras, that is $V_1 \subseteq V_2$ implies that $\mathcal{A}(V_1)$ is a unital C^*-subalgebra of $\mathcal{A}(V_2)$. The *quasilocal algebra* $\mathcal{A}$ is defined to be the inductive limit C^*-algebra of the net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ of local C^*-algebras.

2. *Microcausality* (also called as *Einstein causality*) is the requirement that $\mathcal{A}(V')' \cap \mathcal{A} \supseteq \mathcal{A}(V), V \in \mathcal{K}$, where primes denote spacelike complement and algebra commutant, respectively.

3. *Spacetime covariance.* Let $\mathcal{P}_{\mathcal{K}}$ be the subgroup of the group $\mathcal{P}$ of geometric symmetries of $\mathcal{M}$ leaving the collection $\mathcal{K}$ invariant. A group homomorphism $\alpha\colon \mathcal{P}_{\mathcal{K}} \to \operatorname{Aut}\mathcal{A}$ is given such that the automorphisms $\alpha_g, g \in \mathcal{P}_{\mathcal{K}}$ of $\mathcal{A}$ act covariantly on the observable net: $\alpha_g(\mathcal{A}(V)) = \mathcal{A}(g \cdot V), V \in \mathcal{K}$.

If the quasilocal algebra $\mathcal{A}$ of the local physical theory is commutative, we speak about a *local classical theory*; if it is noncommutative, we speak about a *local quantum theory*. For local classical theories microcausality fulfills trivially.

A *state* ϕ in a local physical theory is defined as a normalized positive linear functional on the quasilocal observable algebra $\mathcal{A}$. The corresponding GNS representation $\pi_\phi\colon \mathcal{A} \to \mathcal{B}(\mathcal{H}_\phi)$ converts the net of C^*-algebras into a net of C^*-subalgebras of $\mathcal{B}(\mathcal{H}_\phi)$. Closing these subalgebras in the weak topology one arrives at a net of local von Neumann observable algebras: $\mathcal{N}(V) := \pi_\phi(\mathcal{A}(V))'', V \in \mathcal{K}$. Von Neumann

algebras are generated by their projections, which are called *quantum events* since they can be interpreted as 0-1–valued observables. The net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of local von Neumann algebras also obeys isotony, microcausality, and $\mathcal{P}_{\mathcal{K}}$-covariance, hence one can also refer to a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of local *von Neumann algebras* as a local physical theory. Although, the local σ-algebras of classical observable events provided by the projections of the local abelian von Neumann algebras are not the most general σ-algebras, still they provide us a rich enough set of examples for classical theories.

One can introduce a number of important locality and causality concepts into the above formalism. Here we only list them in turn and assert their logical relations; for the motivation of these concepts see (Earman and Valente, 2014).

Local primitive causality. For any globally hyperbolic bounded subspacetime region $V \in \mathcal{K}$, $\mathcal{A}(V'') = \mathcal{A}(V)$.

A local physical theory satisfying local primitive causality also satisfies the following two properties:

Local determinism. For any two states ϕ and ϕ' and for any globally hyperbolic spacetime region $V \in \mathcal{K}$, if $\phi|_{\mathcal{A}(V)} = \phi'|_{\mathcal{A}(V)}$ then $\phi|_{\mathcal{A}(V'')} = \phi'|_{\mathcal{A}(V'')}$.

Stochastic Einstein locality. Let $V_A, V_C \in \mathcal{K}$ such that $V_C \subset J_-(V_A)$ and $V_A \subset V_C''$. If $\phi|_{\mathcal{A}(V_C)} = \phi'|_{\mathcal{A}(V_C)}$ holds for any two states ϕ and ϕ' on $\mathcal{A}$ then $\phi(A) = \phi'(A)$ for any projection $A \in \mathcal{A}(V_A)$.

If a net satisfies Haag duality:

$$\mathcal{A}(V')' \cap \mathcal{A} = \mathcal{A}(V) \tag{1}$$

for all bounded globally hyperbolic subspacetime region V, which is a stronger requirement than microcausality, then it also satisfies local primitive causality. But microcausality alone does not entail local primitive causality.

A global version of local primitive causality (entailed by the local one) is

Primitive causality. Let $\mathcal{K}(\mathcal{C}) \subseteq \mathcal{K}$ be a covering collection of a Cauchy surface $\mathcal{C}$ and let $\mathcal{A}(\mathcal{K}(\mathcal{C}))$ be the corresponding algebra. Then $\mathcal{A}(\mathcal{K}(\mathcal{C})) = \mathcal{A}$.

A local physical theory with primitive causality satisfies

Determinism. If $\phi|_{\mathcal{A}(\mathcal{K}_{\mathcal{C}})} = \phi'|_{\mathcal{A}(\mathcal{K}_{\mathcal{C}})}$ for any two states ϕ and ϕ' on $\mathcal{A}$ then $\phi = \phi'$.

In the rest of the paper a local physical theory obeys only isotony, microcausality, and $\mathcal{P}_{\mathcal{K}}$-covariance by definition without any other locality and causality constraints. We turn now to Bell's notion of local causality.

3 Bell's notion of local causality in a local physical theory

Local causality has been playing a central notion in Bell's influential writings on the foundations of quantum theory. To our knowledge it gets an explicit formulation three times: in (Bell, 1975/2004 , p. 54), (Bell, 1986/2004, p. 200), and (Bell, 1990/2004, p. 239-240). In this latter posthumously published paper "La nouvelle cuisine", for example, local causality is formulated as follows:[1]

> "A theory will be said to be locally causal if the probabilities attached to values of local beables in a space-time region V_A are unaltered by specification of values of local beables in a space-like separated region V_B, when what happens in the backward light cone of V_A is already

[1] For the sake of uniformity we slightly changed Bell's denotation and figures.

> sufficiently specified, for example by a full specification of local beables in a space-time region V_C. " (Bell, 1990/2004, p. 239-240)

(For a reproduction of the figure Bell is attaching to this formulation see Fig. 1 with Bell's caption.) Bell

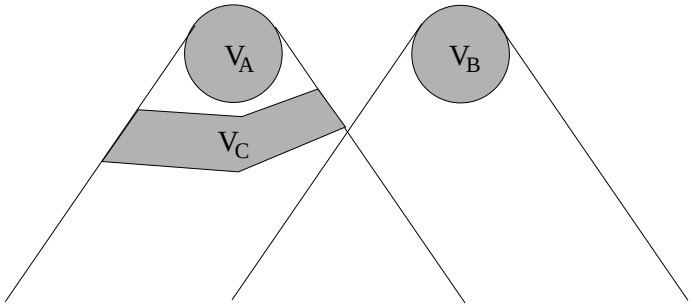


Figure 1: Full specification of what happens in V_C makes events in V_B irrelevant for predictions about V_A in a locally causal theory.

elaborates on his formulation as follows:

> "It is important that region V_C completely shields off from V_A the overlap of the backward light cones of V_A and V_B. And it is important that events in V_C be specified completely. Otherwise the traces in region V_B of causes of events in V_A could well supplement whatever else was being used for calculating probabilities about V_A. The hypothesis is that any such information about V_B becomes redundant when V_C is specified completely." (Bell, 1990/2004, p. 240)

The notions featuring in Bell's formulation has been target of intensive discussion in philosophy of science. Here we would like to give only a brief exposé of them.

The notion "beable" is Bell's neologism. (See Norsen 2009, 2011.) "The *be*ables of the theory are those entities in it which are, at least tentatively, to be taken seriously, as corresponding to something real" (Bell, 1990/2004, p. 234). The clarification of the "beables" of a given theory is indispensable in order to define local causality since "there *are* things which do go faster than light. British sovereignty is the classical example. When the Queen dies in London (long may it be delayed) the Prince of Wales, lecturing on modern architecture in Australia, becomes instantaneously King" (p. 236).

Beables are to be local: "*Local* beables are those which are definitely associated with particular space-time regions. The electric and magnetic fields of classical electromagnetism, $\mathbf{E}(t, x)$ and $\mathbf{B}(t, x)$ are again examples." (p. 234). Furthermore, local beables are to "specify completely" region V_C in order to block causal influences arriving at V_A from the common past of V_A and V_B. (For the question of complete *vs.* sufficient specification see (Seevinck and Uffink, 2014).)

One can translate Bell's above terms in the following way. In a classical field theory beables are characterized by sets of field configurations. Taking the equivalence classes of those field configurations which have the same field values on a given spacetime region one can generate local σ-algebras. Translating σ-algebras into the language of abelian von Neumann algebras one can capture Bell's notion of "local beables" in the framework of a local physical theory. More generally, one can use the term "local beables" both for abelian and also for non-abelian local von Neumann algebras, hence treating local classical and quantum theories on an equal footing.

How to translate the term "complete specification"? Complete specification of field configurations in a given spacetime region means that one specifies the field values to a prescribed value in the given spacetime region, that is one specifies the corresponding local equivalence class of a single configuration. In probabilistic language complete specification is translated into a probability measure having support

on this local equivalence class of the single specified configuration. In the abelian von Neumann language this corresponds to a change of the original state that results in a *pure state* on the local von Neumann algebra in question with value 1 on the projection corresponding to the local equivalence class of the single specified configuration. We also would like this change of states to be as *local* as possible. Both pureness and locality can be captured in a general local physical theory by some conditions imposed on a completely positive map generating the change of states. If the local algebras of the net are *atomic* (which, by the way, is not the case in a general AQFT), the change of states can be generated by conditioning the original state on an arbitrary *atomic* event (a minimal projection) in the local algebra. In this case "complete specification of beables" will mean a so-called selective measurement by an atomic event in a local algebra (Henson, 2013). With these notions in hand we can formulate Bell's notion of local causality in local physical theories:[2]

Definition 2. A local physical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of von Neumann algebras is called *(Bell) locally causal*, if for any pair $A \in \mathcal{N}(V_A)$ and $B \in \mathcal{N}(V_B)$ of projections supported in spacelike separated regions $V_A, V_B \in \mathcal{K}$ and for every locally normal and faithful state ϕ establishing a correlation, $\phi(AB) \neq \phi(A)\phi(B)$, between A and B, and for any spacetime region V_C such that

(i) $V_C \subset J_-(V_A)$,

(ii) $V_A \subset V_C''$,

(iii) $J_-(V_A) \cap J_-(V_B) \cap \big(J_+(V_C) \setminus V_C\big) = \emptyset$,

(see Fig. 2) and for any atomic event C_k of $\mathcal{A}(V_C)$ $(k \in K)$, the following holds:

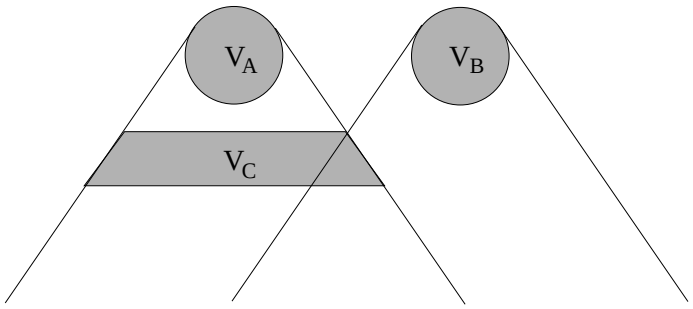


Figure 2: A region V_C satisfying Requirements (i)-(iii).

$$\frac{\phi(C_k A B C_k)}{\phi(C_k)} = \frac{\phi(C_k A C_k)}{\phi(C_k)} \frac{\phi(C_k B C_k)}{\phi(C_k)} \tag{2}$$

Remarks:

1. Again we stress that Definition 2 captures local causality only for local physical theories with *atomic* local von Neumann algebras.

2. In case of classical theories a locally faithful state ϕ determines a locally nonzero probability measure p by $p(A) := \phi(A) > 0, A \in \mathcal{P}(\mathcal{N}(V))$. By means of this (2) can be written in the following 'symmetric' form:

$$p(AB|C_k) = p(A|C_k)p(B|C_k) \tag{3}$$

[2] For a similar approach to local causality using σ-algebras see (Henson, 2013); for a comparison of the two approaches see our (Hofer-Szabó and Vecsernyés 2014).

or equivalent in the 'asymmetric' form:

$$p(A|BC_k) = p(A|C_k) \tag{4}$$

sometimes used in the literature (for example in (Bell, 1975/2004 , p. 54)).

3. The role of Requirement (iii) in the definition is to ensure that "V_C shields off from V_A the overlap of the backward light cones of V_A and V_B". Namely, a spacetime region *above* V_C in the common past of the correlating events (see Fig. 3) may contain stochastic events which, though completely

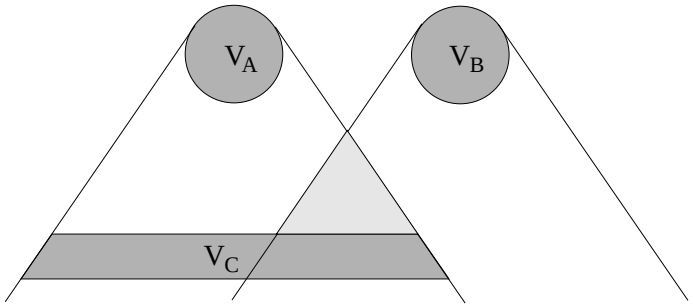


Figure 3: A region V_C for which Requirement (iii) does not hold.

specified by the region V_C, still, being stochastic, could establish a correlation between A and B in a classical stochastic theory (Norsen, 2011; Seevinck and Uffink 2011). If V_C is a piece of a Cauchy surface Requirement (iii) coincides with Requirement (iv):

(iv) $J_-(V_A) \cap J_-(V_B) \cap V_C = \emptyset$

visualized in Fig. 4. However, for algebras corresponding to coverings of Cauchy surfaces Require-

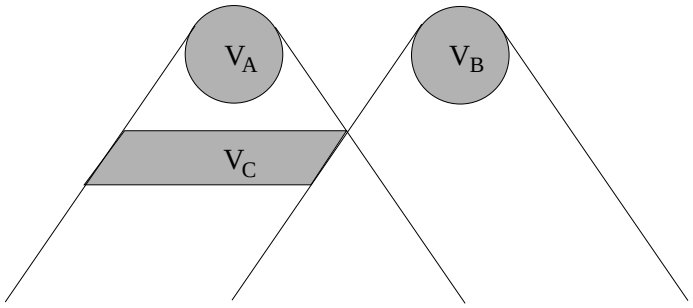


Figure 4: A region V_C for which Requirement (iv) holds.

ment (iii) is weaker than Requirement (iv) since it allows for regions penetrating into the top part of the common past. For local classical theories Requirement (iii) is enough, but for local quantum theories Requirement (iv) should be used.

Of course the main question is how to ensure that a local physical theory is locally causal. Generally the question is difficult to answer; here we simply mention a sufficient condition in case of *atomic* local algebras:

1. A local *classical* theory is locally causal if the local von Neumann algebras are *atomic* and satisfy *local primitive causality*.

   Proof. Due to isotony and local primitive causality $\mathcal{N}(V_A) \subset \mathcal{N}(V_C'') = \mathcal{N}(V_C)$ and hence for any atom C_k of $\mathcal{N}(V_C)$: either (i) $AC_k = 0$ or (ii) $AC_k = C_k$. In case of (i) both sides of (2) is zero, in case of (ii) (2) holds as follows:

$$\frac{\phi(ABC_k)}{\phi(C_k)} = \frac{\phi(BC_k)}{\phi(C_k)} = \frac{\phi(AC_k)}{\phi(C_k)}\frac{\phi(BC_k)}{\phi(C_k)}. \tag{5}$$

2. A local *quantum* theory is locally causal if the local von Neumann algebras are *atomic* and satisfy *local primitive causality*, and if Requirement (iii) in the definition of local causality is replaced by Requirement (iv).

   Proof. Since region V_C is spatially separated from region V_B, $B \in \mathcal{N}(V_B)$ and an atomic event $C_k \in \mathcal{N}(V_C)$ will commute due to microcausality. Using $C_kAC_k = r\,C_k$ (where $r \in [0,1]$ depends on both A and C_k) we obtain:

$$\frac{\phi(C_kABC_k)}{\phi(C_k)} = \frac{\phi(C_kAC_kB)}{\phi(C_k)} = r\frac{\phi(C_kB)}{\phi(C_k)} = \frac{\phi(C_kAC_k)}{\phi(C_k)}\frac{\phi(BC_k)}{\phi(C_k)}. \tag{6}$$

Looking at Point 2 the reader may justly ask: how can a local *quantum* theory be locally causal if local causality implies various Bell inequalities, which are known to be violated for certain set of quantum correlations. Does Definition 2 correctly grasp Bell's intuition of local causality? We answer these questions in the next section.

4 Local causality, Common Cause Principle and the Bell inequalities

Local causality is closely related to Reichenbach's (1956) Common Cause Principle. The *Common Cause Principle* (CCP) states that if there is a correlation between two events A and B and there is no direct causal (or logical) connection between the correlating events, then there always exists a common cause C of the correlation. Reichenbach's original classical probabilistic definition of the common cause can readily be generalized to the local physical theory framework. (See (Rédei 1997, 1998), (Rédei and Summers 2002, 2007), (Hofer-Szabó and Vecsernyés 2012, 2013) and (Hofer-Szabó, Rédei and Szabó 2013).)

Let $\{\mathcal{N}(V), V \in \mathcal{K}\}$ be a net representing a local physical theory. Let $A \in \mathcal{N}(V_A)$ and $B \in \mathcal{N}(V_B)$ be two events (projections) supported in spacelike separated regions $V_A, V_B \in \mathcal{K}$ which correlate in a locally normal and faithful state ϕ. The common cause of the correlation is an event screening off the correlating events from one another and localized in the past of A and B. But in which past? Here one has (at least) three options. One can localize C either (i) in the *union* $J_-(V_A) \cup J_-(V_B)$ or (ii) in the *intersection* $J_-(V_A) \cap J_-(V_B)$ of the causal past of the regions V_A and V_B; or (iii) more restrictively in $\cap_{x \in V_A \cup V_B} J_-(x)$, that is in the spacetime region which lies in the intersection of causal pasts of *every* point of $V_A \cup V_B$. We will refer to the above three pasts in turn as the *weak past*, *common past*, and *strong past* of A and B, respectively (Rédei, Summers, 2007).

Depending on the choice of the past we can define various CCPs in a local physical theory:

Definition 3. A local physical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ is said to satisfy the *(Weak/Strong) CCP*, if for any pair $A \in \mathcal{N}(V_A)$ and $B \in \mathcal{N}(V_B)$ of projections supported in spacelike separated regions $V_A, V_B \in \mathcal{K}$ and for every locally faithful state ϕ establishing a correlation between A and B, there exists a nontrivial common cause system, that is a set of mutually orthogonal projections $\{C_k\}_{k \in K} \subset \mathcal{N}(V_C), V_C \in \mathcal{K}$ summing up to the unit of the algebra, satisfying

$$\frac{\phi(C_kABC_k)}{\phi(C_k)} = \frac{\phi(C_kAC_k)}{\phi(C_k)}\frac{\phi(C_kBC_k)}{\phi(C_k)}, \qquad \text{for all } k \in K \tag{7}$$

such that the localization region of V_C is in the (weak/strong) common past of V_A and V_B.

A common cause is called *nontrivial* if $C_k \not\leq X$ with $X = A, A^\perp, B$ or $B^\perp$ for some $k \in K$. If $\{C_k\}_{k\in K}$ commutes with both A and B, then we call it a *commuting* common cause system, otherwise a *noncommuting* one, and the appropriate CCP a *Commutative/Noncommutative CCP*.

The status of these six different notions of the CCP has been thoroughly scrutinized in a special local quantum theory, namely algebraic quantum field theory (AQFT). Here we only give a brief overview.

The question whether the Commutative CCPs are valid in a Poincaré covariant local quantum theory was first raised by Rédei (1997, 1998). As an answer, Rédei and Summers (2002, 2007) have shown that the Commutative Weak CCP is valid in Poincaré covariant AQFT. Since local algebras in a Poincaré covariant AQFT are atomless (type III) von Neumann algebras, the question has been raised whether Commutative Weak CCP is valid in local quantum theories with locally finite dimensional, hence atomic local von Neumann algebras. Deciding the question, Hofer-Szabó and Vecsernyés (2012a) have given an example in the local quantum Ising model where the Commutative Weak CCP is *not* valid. A natural reaction to these facts was to ask what role commutativity plays in these propositions. Addressing this question, Hofer-Szabó and Vecsernyés (2013) have shown that allowing common causes *not* to commute with the correlating events, the Noncommutative Weak CCP can be proven in local (UHF-type) quantum theories with finite dimensional local von Neumann algebras.

Concerning the Commutative (Strong) CCP less is known. If one also admits projections localized only in *un*bounded regions, then the Strong CCP is known to be false: von Neumann algebras pertaining to complementary wedges contain correlated projections but the strong past of such wedges is empty (see (Summers and Werner, 1988) and (Summers, 1990)). In spacetimes having horizons, e.g. those with Robertson–Walker metric, the common past of spacelike separated bounded regions can be empty, although there are states which provide correlations among local algebras corresponding to these regions (Wald 1992). Hence, CCP is not valid there. Restricting ourselves to projections in *local* algebras on *Minkowski* spacetimes the situation is not clear. We are of the opinion that one cannot decide on the validity of the (Strong) CCP in this case without an explicit reference to the dynamics.

Now, what is the relationship between the various CCPs and Bell's local causality? The following list of *prima facie* similarities and differences may help to explicate this relationship:

Similarities:

1. Both local causality and the CCPs are *properties of a local physical theory* represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$.

2. The core mathematical requirement of both principles is the *screening-off condition* (2) or equivalently (7).

3. The *Bell inequalities* can be derived from both principles. (But see below.)

Differences:

1. In case of local causality the screening-off condition (2) is required for *every* atomic event (satisfying certain localization conditions). In case of the CCP for every correlation *only a single* subset of events is postulated satisfying the screening-off condition (7).

2. In case of local causality the screening-off condition is required only for *atomic* events. In case of the CCPs these atomic screener-offs of the algebra $\mathcal{A}(V_C)$ are called trivial, since they screen any correlation off. What one is typically looking for are *nontrivial* common causes.

3. In case of local causality screener-offs are localized *'asymmetrically'* in the past of V_A; in case of the CCP they are localized *'symmetrically'* in either the weak, common or strong past of V_A and V_B.

Let us come back to Point 1 of the Similarities, that is to the relation of local causality and the CCPs to the Bell inequalities. In (Hofer-Szabó and Vecsernyés, 2013b, Proposition 2) we have proven a proposition which clarifies the relation between the CCPs and the Bell inequalities. It asserts that the Bell inequalities can be derived from the existence of a (local, non-conspiratorial joint) common cause system for a set of correlations *if* common causes are understood as *commuting* common causes. However, if we also allow for *non*commuting common causes, the Bell inequalities can be derived only for *another* state which is *not* identical to the original one. And indeed in (Hofer-Szabó and Vecsernyés, 2013a,b) a noncommuting common cause was constructed for a set of correlations violating the Clauser–Horne inequality. Moreover, this common cause was localized in the *strong* past of the correlating events.

Now, an analogous proposition holds for the relation between local causality and the Bell inequalities. We assert here only the proposition without the proof since the proof is step-by-step the same as that of the proposition mentioned above.

Proposition 1. Let $\{\mathcal{N}(V), V \in \mathcal{K}\}$ be a locally causal local physical theory with atomic (type I) local von Neumann algebras. Let $A_1, A_2 \in \mathcal{A}(V_A)$ and $B_1, B_2 \in \mathcal{A}(V_B)$ be four projections localized in spacelike separated spacetime regions V_A and V_B, respectively, which pairwise correlate in the locally faithful state ϕ that is

$$\phi(A_m B_n) \quad \neq \quad \phi(A_m)\,\phi(B_n) \tag{8}$$

for any $m, n = 1, 2$. Let furthermore $\{C_k\}_{k\in K} \subset \mathcal{N}(V_C), V_C \in \mathcal{K}$ be a maximal partition of the unit, where the set $\{C_k\}_{k\in K}$ contains mutually orthogonal *atomic* projections satisfying Requirements (i)-(iii) in Definition 2 of local causality. Then the Clauser–Horne inequality

$$-1 \leqslant \phi_{\{C_k\}}(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1) \leqslant 0. \tag{9}$$

holds for the state $\phi_{\{C_k\}}(X) := \sum_k \phi(C_k X C_k)$. If $\{C_k\}$ commutes with A_1, A_2, B_1 and B_2, then the Clauser–Horne inequality holds for the original state ϕ:

$$-1 \leqslant \phi(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1) \leqslant 0. \tag{10}$$

The moral is the same as in the case of the CCPs: the Bell inequalities can be derived in a locally causal local physical theory only for a modified state $\phi_{\{C_k\}}$; it can be derived for the original state ϕ *if* the set of atomic projections $\{C_k\}$ localized in V_C commutes with A_1, A_2, B_1 and B_2. What is needed for this to be the case?

In local *classical* theories any element taken from any local algebra will commute, therefore the Bell inequalities will hold in local classical theories. In locally causal local *quantum* theories, commutativity of $\{C_k\}$ and the correlating events is not guaranteed. If V_C is spatially separated from V_B (due to Requirement (iv) in Definition 2), then $\{C_k\}$ will commute with B_1 and B_2 and hence (2) will be satisfied. However, for noncommuting A_1 and A_2 one cannot pick a maximal partition $\{C_k\}$ commuting with both projections, and therefore the theorem of total probability, $\sum_k \phi(C_k A_m C_k) = \phi(A_m)$, will not hold for the original state ϕ at least for one of the projections A_1 and A_2 (it will hold only for the state $\phi_{\{C_k\}}$). This fact blocks the derivation of Bell inequalities for the original state ϕ. (For the details see (Hofer-Szabó and Vecsernyés, 2013b, p. 410) In short, the Bell inequalities can be derived in a locally causal local quantum theory only if all the projections commute.

Coming back to the question posed at the end of the previous Section, namely how a local *quantum* theory can be locally causal in the face of the Bell inequalities, we already know the answer: the Bell inequalities can be derived from local causality if it is required that the 'beables' of the local theory are represented by *commutative* local algebras. This fact is completely analogous to the relation shown in (Hofer-Szabó and Vecsernyés, 2013b), namely that the Bell inequalities can be derived from a (local, non-conpiratorial, joint) common cause system if it is a *commuting* common cause system. Thus, the

violation of the Bell inequalities for certain quantum correlations is compatible with locally causal local quantum theories but not with locally causal local classical theories. Local causality is a more general notion than captured by the Bell inequalities.

5 Conclusions

In this paper we have shown the following:

(i) Bell's notion of local causality presupposes a clear-cut framework in which probabilistic and spatiotemporal entities can be related. This aim can be reached by introducing the notion of a *local physical theory* represented by an isotone net of algebras.

(ii) Within this general framework we have defined Bell's notion of *local causality* and shown sufficient conditions on which local physical theories will be locally causal.

(iii) Finally, we pointed out some important similarities and differences between local causality and the CCPs and showed that in a locally causal local quantum theory one cannot derive the Bell inequalities from local causality just as one cannot derive them from noncommuting common causes.

Acknowledgements. This work has been supported by the Hungarian Scientific Research Fund OTKA K-100715 and K-108384.

References

J.S. Bell, *Speakable and Unspeakable in Quantum Mechanics*, Cambridge: Cambridge University Press, (2004).

J. Butterfield, "Vacuum correlations and outcome independence in algebraic quantum field theory" in D. Greenberger and A. Zeilinger (eds.), *Fundamental Problems in Quantum Theory, Annals of the New York Academy of Sciences, Proceedings of a conference in honour of John Wheeler*, 768-785 (1995).

J. Butterfield, "Stochastic Einstein Locality Revisited," *British Journal for The Philosophy of Science*, **58**, 805-867, (2007).

J. Earman and G. Valente, "Relativistic causality in algebraic quantum field theory," (manuscript), (2014).

R. Haag, *Local Quantum Physics*, (Springer Verlag, Berlin, 1992).

H. Halvorson, "Algebraic quantum field theory," in J. Butterfield, J. Earman (eds.), *Philosophy of Physics, Vol. I*, Elsevier, Amsterdam, 731-922 (2007).

J. Henson, "Non-separability does not relieve the problem of Bell's theorem," *Found. Phys.*, **43**, 1008-1038 (2013).

G. Hofer-Szabó, M. Rédei and L. E. Szabó, *The Principle of the Common Cause*, Cambridge: Cambridge University Press, 2013

G. Hofer-Szabó and P. Vecsernyés, "Reichenbach's Common Cause Principle in AQFT with locally finite degrees of freedom," *Found. Phys.*, **42**, 241-255 (2012a).

G. Hofer-Szabó and P. Vecsernyés, "Noncommuting local common causes for correlations violating the Clauser–Horne inequality," *Journal of Mathematical Physics*, **53**, 12230 (2012b).

G. Hofer-Szabó and P. Vecsernyés, "Noncommutative Common Cause Principles in AQFT," *Journal of Mathematical Physics*, **54**, 042301 (2013a).

G. Hofer-Szabó and P. Vecsernyés, "Bell inequality and common causal explanation in algebraic quantum field theory," *Studies in the History and Philosophy of Modern Physics*, **44 (4)**, 404-416 (2013b).

G. Hofer-Szabó and P. Vecsernyés, "On Bell's local causality in local classical and quantum theory," submitted (2014).

T. Norsen, "Local causality and Completeness: Bell vs. Jarrett," *Found. Phys.*, **39**, 273 (2009).

T. Norsen, "J.S. Bell's concept of local causality," *Am. J. Phys*, **79**, 12, (2011).

M. Rédei, "Reichenbach's Common Cause Principle and quantum field theory," *Found. Phys.*, **27**, 1309-1321 (1997).

M. Rédei and J. S. Summers, "Local primitive causality and the Common Cause Principle in quantum field theory," *Found. Phys.*, **32**, 335-355 (2002).

H. Reichenbach, *The Direction of Time*, (University of California Press, Los Angeles, 1956).

M. P. Seevinck and J. Uffink, "Not throwing out the baby with the bathwater: Bell's condition of local causality mathematically 'sharp and clean', " in: Dieks, D.; Gonzalez, W.J.; Hartmann, S.; Uebel, Th.; Weber, M. (eds.) *Explanation, Prediction, and Confirmation The Philosophy of Science in a European Perspective*, Volume 2, 425-450 (2011).

S. J. Summers and R. Werner, "Bell's inequalities and quantum field theory, I: General setting," *Journal of Mathematical Physics*, **28**, 2440-2447 (1987a).

S. J. Summers and R. Werner, "Bell's inequalities and quantum field theory, II: Bell's inequalities are maximally violated in the vacuum," *Journal of Mathematical Physics*, **28**, 2448-2456 (1987b).

S. J. Summers and R. Werner, "Maximal violation of Bell's inequalities for algebras of observables in tangent spacetime regions," *Ann. Inst. Henri Poincaré – Phys. Théor.*, **49**, 215-243 (1988).

DERIVING GENERAL RELATIVITY FROM STRING THEORY

NICK HUGGETT, UNIVERSITY OF ILLINOIS AT CHICAGO
AND
TIZIANA VISTARINI, RUTGERS UNIVERSITY

1. Introduction

The goal of this paper is to explain the significance of the conformal symmetry of string theory. Along the way we will introduce the basics of string theory in a streamlined fashion, drawing on familiar ideas from classical and quantum field theory. We will then explain how general relativity is a 'consequence' of string theory: not merely in the sense that it contains massless spin-2 particles – gravitons – but in the very strong sense that the coherent states of the graviton obey the Einstein field equations – gravitons truly form the gravitational field. This result follows from reimposing conformal symmetry in *quantized* string theory; so in the final section of the paper we sketch some more esoteric considerations justifying this assumption.

2. The Formalism

2.1. **The Classical String.** We will start with a classical relativistic string[1], an object of one spatial dimension and one temporal dimension – it's best to think of it as a spacetime object from the get go. Let us suppose that it is 'closed', meaning that its ends are joined into a loop (figure 2.1 shows an open string – to close it, the timelike edges should be identified). Our string is free, subject to internal tension, but (for now) under the influence of no external forces, including gravity, so that it lives in Minkowski spacetime, with metric $\eta_{\mu\nu}$. Suppose that the points of the string come labelled with 'internal spacetime coordinates' τ and σ (later σ_0 and σ_1); while 'external' or 'target' spacetime has coordinates X^μ ($\mu = 0, 1, \ldots, D-1$). Then we can describe the string worldsheet in spacetime by assigning appropriate coordinates $(X^0, X^1, \ldots, X^{D-1})$ to each internal point (σ, τ); formally there is a D-component vector *field* on the string. From the point of view of the string then, motion in target space amounts to changes in this field. This picture will be important as we progress, so bear it in mind.

So how do we expect this 2-dimensional object to behave? One's mind turns to Hooke's law, but that is uncongenial to relativity – Lorentz contraction should not change the tension in a string. What Hooke's law tells us more generally is that a string will minimize its length: again, not relativistically invariant, but close – the relativistic statement is that a string will minimize its *spacetime* area. Thus the simplest classical, relativistic string

[1]Here we draw heavily on several recent text-books, especially [Becker et al. (2006), Kiritsis (2011), Polchinski (2003), Zwiebach (2004)].

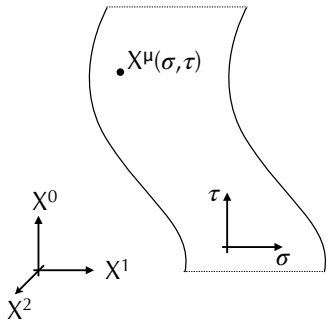


FIGURE 1. An open string in target space – if the timelike edges are identified then it becomes a closed string.

action is proportional to the invariant area $S = -T \int \mathrm{d}A$. Explicitly, $\mathrm{d}A = \sqrt{-g} \cdot \mathrm{d}X^\mu \mathrm{d}X^\nu$, or transforming into string coordinates, we obtain the famous Nabu-Goto action:

$$S_{NG} = -T \int \mathrm{d}\sigma^2 \sqrt{-\det\Big(\eta_{\mu\nu} \frac{\partial X^\mu}{\partial \sigma^\alpha} \frac{\partial X^\nu}{\partial \sigma^\beta}\Big)}. \tag{1}$$

T is the tension in the string (though you can't immediately see this from the form of the action); it makes clear that the string does not satisfy Hooke's law, because it is an invariant *constant*. The action also shows that all that matters is the total length of the string, not how parts might be stretched relative to one another – again un-Hooke-like behaviour. So, for one thing, the dynamics has no way of identifying parts of the string over time (the action has diffeomorphism symmetry with respect to the σs) – but the significance of this behaviour is far greater.[2]

The square root in S_{NG} is awkward, but a formal trick leads to the equivalent sigma (or Polyakov) action:

$$S_\sigma = -\frac{T}{2} \int \mathrm{d}^2\sigma \sqrt{-\gamma} \gamma^{\alpha\beta} \eta_{\mu\nu} \frac{\partial X^\mu}{\partial \sigma^\alpha} \frac{\partial X^\nu}{\partial \sigma^\beta}. \tag{2}$$

The 'trick' involves introducing a second 'internal' metric, $\gamma_{\alpha\beta}$ on the string worldsheet – to be carefully distinguished from the restriction of the spacetime metric to the string.

Now the un-Hooke-like behaviour of the string manifests itself in the fact that intervals with respect to the internal metric have no physical significance. The action appears to depend on how the string is stretched along its length – the derivatives are determined by the distance in external space separating points on the worldsheet separated by an infinitesimal distance in the string coordinates. But the behavior of the string that we have been stressing means that such infinitesimal distances have no physical significance,

[2]Quick aside: in fact one can give a Hooke's law treatment of the string, not in inertial coordinates, but in 'light cone' coordinates, in which one spatial coordinate is 'boosted to the speed of light'. Such coordinates are used in most text-books at some point.

and so it should make no difference if any is rescaled by an arbitrary factor. In short, the action must be Weyl, or conformally invariant, and indeed it is: as can be readily checked, the sigma action is unchanged by $\gamma_{\alpha\beta} \rightarrow e^{\omega(\tau,\sigma)}\gamma_{\alpha\beta}$.[3]

A couple of short notes. First, the action is conformally invariant with respect to the string metric, not the target space metric! For η the relevant symmetry is Poincaré invariance. (The other symmetry of the action is diffeomorphism invariance with respect to both the σs and Xs.) Second, although we have been stressing the connection of conformal invariance to the un-Hooke-like behavior of a relativistic string, we did so mainly to illustrate how string theory is grounded in some very familiar physics. Conformal invariance will be crucial in what follows, but all we need is the straight-forward mathematical fact that S_σ has that symmetry – not any story about why. For now we have the following: from the point of view of the string, string theory concerns a D-dimensional conformal field, living on a 2-dimensional spacetime (i.e., the worldsheet). That picture was central to the developments of the 'second string revolution' of the 1990s, and generally is the one that we will adopt.

The symmetries can be used to set the worldsheet metric flat:

$$S_\sigma = \frac{T}{2}\int \mathrm{d}^2\sigma \dot{X}^2 - X'^2, \tag{3}$$

where the derivative are with respect to the worldsheet coordinates.[4] The corresponding Hamiltonian is:

$$H = \frac{T}{2}\int d\sigma\ \dot{X}^2 + X'^2; \tag{4}$$

and minimizing with respect to X^μ yields a wave equation,

$$\ddot{X}^\mu - X''^\mu = 0. \tag{5}$$

The general solution for a closed string is (after a little more work in classical wave physics):

$$X^\mu = X_0^\mu + \ell_s^2 p^\mu \tau + i\frac{\ell_s}{\sqrt{2}}\sum_{n\neq 0}\frac{1}{n}\big(\alpha_n^\mu e^{-i2n(\tau-\sigma)} + \bar{\alpha}_n^\mu e^{-i2n(\tau+\sigma)}\big), \tag{6}$$

where ℓ_S, the 'characteristic string length', is determined by the tension: $\ell_s^2 = 1/T$. This equation describes an initial position, linear momentum, and left- and right-moving vibrations – the α_n are the amplitudes of the modes of the string. Identifying the linear momentum as the zeroth mode of the string will be useful.

[3]More carefully, we have been talking about Weyl symmetry; 'conformal' transformations are strictly a sub-group of the diffeomorphisms, namely those whose only effect is to introduce a Weyl factor. This point is, for instance, important for understanding why conformal symmetry remains in (3), even though the Weyl symmetry has been gauge fixed.

[4]In the following the reader is especially referred to [Becker et al. (2006), §2.2-3]

$$\alpha_0^\mu \equiv \frac{\ell_s}{2} p^\mu \equiv \bar{\alpha}_0^\mu. \tag{7}$$

Substituting the mode expansion of X^μ into the Hamiltonian (4) gives

$$H = \sum_{n=-\infty}^{\infty} (\alpha_{-n} \cdot \alpha_n + \tilde{\alpha}_{-n} \cdot \tilde{\alpha}_n) = 0. \tag{8}$$

2.2. **Immediate Consequences.** Now, because of conformal symmetry, the variation of the action with respect to rescaling the metric must vanish:

$$0 = \frac{1}{\sqrt{-h}} \frac{\delta S_\sigma}{\delta h^{\alpha\beta}} = -2T\pi\big(\partial_\alpha X \cdot \partial_\beta X + \frac{1}{2} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}_{\alpha\beta} (\dot{X}^2 - X'^2)\big). \tag{9}$$

The four equations given by the possible values of α and β can be solved, and if the expansion for X^μ is inserted, entail that:

$$\forall m \in \mathbb{Z} \quad 0 = \frac{1}{2} \sum_{n=-\infty}^{\infty} \alpha_{m-n} \cdot \alpha_n \equiv L_m. \tag{10}$$

The L_m and $\tilde{L}_m$ are crucial objects in the formalism describing the string, the '*Virasoro generators*', and the constraints (10) play a vital role in the theory. Physically speaking, from the worldsheet perspective, (9) gives $T_{\alpha\beta}$, the stress-energy tensor of the 2-dimensional stringy spacetime, and the Virasoro generators are its modes. Geometrically speaking, they are the generators of conformal transformations on the worldsheet.

As an example, consider the role of the constraints in determining the mass spectrum of the string. Observed at scales well above its characteristic length, intuitively a string will appear as a (spatially) point-like object – a particle – since its extension 'can't be seen'. Since the string appears as a particle, its linear four-momentum must satisfy the usual relation to its rest mass. Using (7):

$$-M^2 = p^2 = \frac{2(\alpha_0^2 + \tilde{\alpha}_0^2)}{\ell_s^2}. \tag{11}$$

But the $m = 0$ Virasoro constraint yields

$$L_0 = \frac{1}{2} \sum_{n=-\infty}^{\infty} \alpha_{-n} \cdot \alpha_n = 0 \quad \Rightarrow \quad \frac{\alpha_0^2}{2} = -\sum_{n=1}^{\infty} \alpha_{-n} \cdot \alpha_n, \tag{12}$$

and similarly for $\tilde{L}_0$. So, using (8)

$$-M^2 = \frac{4}{\ell_s^2}\sum_{n=1}^{\infty}(\alpha_{-n}\cdot\alpha_n + \tilde{\alpha}_{-n}\cdot\tilde{\alpha}_n) = \frac{4}{\ell_s^2}H. \tag{13}$$

In other words, it follows from the constraint that the 'particle-mass' of a string depends on its vibrational modes – different vibrations give different 'particles'. Moreover, the Hamiltonian is proportional to the mass *squared* of an excited string, not (as might have been expected in relativity) the mass.

2.3. **Quantization.** In this paper we employ both canonical and path integral quantization: either way, X^μ is a field on a 2-dimensional Minkowski spacetime – the string worldsheet. Thus we start with equal-time commutation relations on the 'field':

$$[X^\mu(\sigma), \Pi^\nu(\sigma')] = i\eta^{\mu\nu}\delta(\sigma - \sigma'), \tag{14}$$

which entails via 6 that

$$[\alpha_m^\mu, \alpha_n^\nu] = m\eta^{\mu\nu}\delta_{m+n}. \tag{15}$$

Hence the quantized αs are raising and lowering operators, as one should expect. Our earlier analysis now shows that the quantized mass spectrum is discrete: it includes massless photons and gravitons, and importantly, for later work, a new masless scalar, the 'dilaton' (as well as negative mass tachyon modes). For suitable string tensions, the modes can reproduce the observed masses of meson families. However, the appropriate tension for quantum gravity is much greater, so observed particles are *not* theorized to be mode excitations of the string. (The mass spectrum of the standard model, is reproduced in a more complex way – relying, for instance, on compactified dimensions or D-branes.)

(15) tells us that for $m \neq 0$ the Virasoro generators can be obtained by simply replacing the αs in (10) with operators; while L_0 requires normal ordering. Omiting all details, the resulting commutation relations are found to be:

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{D}{12}m(m^2-1)\delta_{m+n,0}, \tag{16}$$

the (classical) algebra of conformal generators, plus a 'central charge' term, which indicates a quantum 'anomaly', a breakdown of classical conformal invariance. Restoring the symmetry requires $D = 26$ and leads to the infamous compactified dimensions of string theory. In what follows we explain another consequence of the anomaly.

3. General Relativity from String Theory

Consider the sigma-action, (2) but with a general Lorentzian metric:

$$S_\sigma = -\frac{1}{\alpha'}\int \mathrm{d}^2\sigma\ \sqrt{-\gamma}\gamma^{\alpha\beta}G_{\mu\nu}\partial_\alpha X^\mu \partial_\beta X^\nu. \tag{17}$$

α' is (up to a factor) the reciprocal of the tension – in worldsheet perturbation theory, an expansion parameter. Otherwise, the only change is $\eta_{\mu\nu} \to G_{\mu\nu}$. At this point you may think that G is free parameter, to be inserted by hand – that the 'background' metric is independent of what the strings do. But you would be wrong – G has to satisfy the source-free Einstein field equations (a result going back to [Friedan (1980)]; and conjectures going back to the 1970s). *String theory requires general relativity (to lowest order).*

The proof runs as follows ([Callan et al. (1985)], see [Gasperini (2007)] for more detail):

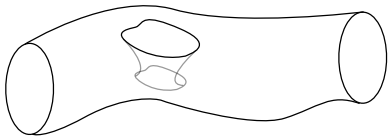


FIGURE 2. A string worldsheet with a 'hole' through it. This surface represents the first order correction to the closed string propagator – time is running left-to-right.

(1) The perturbative expansion of the interacting string propagator is a sum of tori – with two legs – of increasing genus. Consider the first-order term in figure 3.
(2) Its contribution comes from a QFT defined by (17): from the point of view of the string, a 2-dimensional QFT – a 'non-linear sigma model' – in which $G_{\mu\nu}$ is a (varying) self-coupling for the X^μ field.
(3) This QFT has a well-studied perturbation theory, and is known to be renormalizable. (Here we have second perturbation expansion: one on the worldsheet, which is itself a term in the expansion of the string propagator.)
(4) Renormalization means counter-terms, which means running terms, with a length dependence. The renomalization group studies this dependence, describing the behavior in terms of $\beta-$functions.
(5) Friedan studied this renormalization group behavior, showing perturbatively that the β-function associated with $G_{\mu\nu}$ is – remarkably – given by $\beta_G = R_{\mu\nu} + O(\alpha')$, the Ricci tensor plus higher order corrections.
(6) Now, recall that we are talking about a field theory on a string worldsheet, so the length scale is with respect to the string metric, rather than the target space metric; but conformal invariance means that the theory can have no dependence on such a length! So the β-function vanishes, and by Friedan's work, $R_{\mu\nu} = 0$, or $G_{\mu\nu}$ is 'Ricci flat' to lowest order. But that's equivalent to the vacuum Einstein field equations.

In short, the conformal invariance of string theory as a worldsheet QFT entails general relativity for target space. Of course, spacetime is not Ricci flat, but happily the result generalizes. To take a particularly salient example, suppose that strings live in a target space with a Yang-Mills field, a string theory with familiar matter. The action has the form:

$$S_\sigma \sim \frac{1}{\alpha'} \int \mathrm{d}^2\sigma \mathrm{d}\theta \; G_{\mu\nu}\partial X^\mu \partial X^\nu + A_{\mu a}(X)\partial X^\mu j^a + \theta\psi^i\partial\psi^i. \tag{18}$$

This action is for a 'heterotic' string, featuring both conformal symmetry and supersymmetry between bose and fermi degrees of freedom: θ represents fermionic 'coordinates' in addition to the bosonic σs. A is the background gauge field, and ψ the fermions to which it couples (j their current). [Callan et al. (1985)] investigated this action, showing that to lowest order $G_{\mu\nu}$ is Ricci flat, and that the standard free Yang-Mills equations must be satisfied. Moreover, they also showed that at first order in α' the β-function for the target metric has a term corresponding to the stress-energy of the Yang-Mills field.

$$\beta_G = R_{\mu\nu} - \frac{\alpha'}{2}\alpha'\mathrm{tr}(F^2_{\mu\nu}) + O(\alpha') \text{ terms for the coupling to } G_{\mu\nu}. \tag{19}$$

Thus, once again, worldsheet conformal symmetry, $\beta_G = 0$, means that to order α', the EFEs hold, even if when matter is present. However, we want to argue that these are more than formal results (we express what seems to be the view of string theorists). For the results do not simply show that strings tell matter and geometry backgrounds how to behave: they in fact describe how fields built from string excitations are related, so that the field equations are simply low energy descriptions of the string itself.

The reason is that 'background' fields do not represent new degrees of freedom in addition to those of string theory: they are not distinct primitive entities. Instead they represent the behavior of coherent states of string excitations: the quantum states, that is, which describe classical field behavior. Thus when one includes a general metric in the action, one has a quantum theory of perturbations around a coherent state corresponding to the given classical metric. The result shows that there is not a free choice of background fields, but that graviton and – in this case – Yang-Mills quanta coherent states must be related appropriately: by the EFEs.

More precisely, there are two claims involved in the view that background fields represent coherent states (both with evidence in their favor, [**?**, §3.6]): (i) that the string is an adequate 'theory of everything', in the sense that the string spectrum includes quanta for all desired background fields and (ii) that the terms in the action accurately capture the effective behavior of those coherent states. Then by (i) the $G_{\mu\nu}$ field is composed of stringy excitations, and by (ii) it satisfies the EFEs, making it the gravitational field, and the excitations gravitons. Thus, in the most literal sense, the general relativistic theory of spacetime is a low energy effective theory of strings.

If this is correct, then in a central sense, string theory is background independent: the metric arises from string interactions, rather than being stipulated a priori. Just like general relativity, many solutions are possible, but matter and gravity have to satisfy a mutual dynamics – except in string theory, there is no fundamental distinction between the two, a significant ontological unification.

What is left of the charge of background dependence? One might take the view that since the derivation starts with (2), which manifestly involves a Minkowski metric, at least that much geometry is 'background' (even if interpreted as an inner product on X^μ fields).

It would still follow that the really interesting geometry of the theory is dynamical, and just an aspect of the same processes that constitute matter. But the situation is even better. For addressing the question of background independence, one should instead start with the more general action (17). The results above then show that the only possible metrics are Ricci flat, and so the only 'background' assumptions involves selecting one such metric. The idea that the geometry has to be 'put in by hand' hardly applies at all.[5].

4. Towards a Philosophical Analysis of Conformal Symmetry

The connections to the dimensionality of spacetime, and to the EFEs that we have sketched show that conformal symmetry is a key concept connecting string theory to phenomenological spacetime. (By 'phenomenological' we mean space and time as they are described by general relativity; and thus as they are envisioned in more-or-less direct experience, since the situations we normally experience fall within the domain of general relativity.) We claim that *conformal symmetry should therefore be an important focus of philosophical attention in the study of string theory.* [Huggett and Wüthrich (2013)] argues for the importance of philosophical analysis of the 'empirical significance' of such concepts in theories of quantum gravity: in short, such analyses promise to illuminate how (and even whether) aspects of spacetime can emerge from a theory which does not presuppose them at the fundamental level, in some sense. (The paper also explains the value of such analyses even for partial theories, such as bosonic string theory: the development of complete theories must be *preceded* by the development of suitable concepts, typically found in proto-form in incomplete theories.) This essay is a contribution to such an analysis, which we continue to develop in this section. The key question now is whether conformal symmetry is an independent postulate of string theory. We suggest not: the above results do not require extra assumptions, but are essential to string theory.

The story so far is that the classical string action is conformally invariant, but that this symmetry is broken by quantization. We have seen the effect of this 'conformal anomaly' in the central charge appearing in the Virasoro algebra, and in slightly more detail in the derivation of the EFEs, above.[6]

[5]Jeffrey Harvey has noted that even classical relativity requires some background, say in the form of asymptotic behaviour

[6]Now, the latter derivation was given in the context of renormalization, and so may appear to be a consequence of perturbation theory; one might assume that string theory does not require general relativity intrinsically, but only in order for a certain kind of approximation scheme to work. This would be a mistake (see e.g., [Nakahara (2003)]). The short story is this. Consider a schematic path integral for a quantum field:

$$\int \mathcal{D}\varphi \; e^{i \int \mathbf{d}x \; \mathcal{L}[\varphi]}. \tag{20}$$

Manifestly, invariance of the Langrangian under a symmetry no longer suffices for a quantum symmetry: path integrals – hence amplitudes – will only be unchanged if the measure of the path integral is also invariant, and anomalies arise when it is not. Of course anomalies show up in the perturbative expansion of a path integral, but path integrals themselves are not inherently perturbative. Much more could be said on this subject, but not in this place.

At this point one might wonder whether it is possible to abandon conformal symmetry.[7] Thinking through the symmetries of the action (2), what this means is that Weyl transformations are no longer gauge symmetries, only diffeomorphisms are. Of course that would make a nonsense of the development of the string given earlier in this paper – indeed, it's simply false of the action classically! But the point just made is that the conformal anomaly means that the quantum system need not have the same symmetry. In this case of course different choices of conformal factor in the Weyl transformation of the internal metric, $\gamma_{\alpha\beta} \rightarrow e^{\omega(\tau,\sigma)}\gamma_{\alpha\beta}$, will be physically different: hence $\omega(\tau,\sigma)$ is a new physical degree of freedom over the worldsheet, in addition to, and prima facie rather alike, the X^μs. However, $\omega \rightarrow \omega + \lambda$, a 'translation' in this new 'dimension', means a conformal transformation on the world sheet, since $e^{\omega+\lambda} = e^\omega e^\lambda$. Hence, if Weyl symmetry fails, so does translation symmetry in this new 'ω-dimension'. For this reason the ω field cannot be just an additional target spacetime coordinate but requires a different interpretation, as a scalar background field. As a matter of fact, it has the form of a background *dilaton* field, $\Phi(X)$, the new string mode introduced earlier: more specifically (as we shall discuss) it is a *linear* dilaton field.

We shall explore the consequences of the failure of Weyl symmetry by investigating the dilaton – to summarize what we just said, its appearance is a direct consequence of violating the symmetry. On the one hand the dilaton allows one to relax some of the consequences of Weyl symmetry discussed so far. On the other, it will allow us to make good on our claim that Weyl symmetry is not an independent postulate of string theory, in the sense that it will in general signal a breakdown of perturbation theory.

To sketch the physics of the linear dilaton, we start with a new action including a background dilaton field $\Phi(X)$, as usual understood as representing physics around a coherent state (here of the scalar dilaton):

$$\frac{1}{4\pi\alpha'}\int_\Sigma d^2\sigma\sqrt{-\gamma}[(\gamma^{ab}G_{\mu\nu}(X) + \alpha' R\Phi(X)], \tag{21}$$

where R is still the Ricci scalar, and α' is still the expansion parameter of string perturbation theory (the reciprocal of the tension).

As we discussed above, formally (at least) this is the action for a two-dimensional interacting field theory. But the X^μ fields can be re-interpreted as target space coordinates; moreover at low energy it can be rewritten as an effective low energy action over *spacetime*:

$$S_X = \frac{1}{2}\int d^D X\sqrt{-G}e^{-2\Phi}[-\frac{2(D-26)}{3\alpha'} + R + 4\partial_\mu\Phi\partial^\mu\Phi + O(\alpha')]. \tag{22}$$[8]

Comparison of (21) and (22) shows that at low energy the perturbative expansion parameter α' – (a function of) the string coupling – can be identified with $e^{2\Phi(X)}$. This identification indicates the link between the presence of a linear dilaton and the applicability of string

[7] The remainder of the section draws heavily on [Polchinski (2003)] §3.4, 3.7 and 9.9.

[8] [Polchinski (2003), §3.7].

perturbation theory: where the former diverges, the latter breaks down. Below we will indicate how controlling the divergences of the linear dilaton – hence the existence of string perturbation theory – leads back to Weyl invariance; for now, in order to explain some of the features of the linear dilaton, we will simply assume that result. More specifically, we assume that the β-functions for the action (22) vanish:

$$\beta^{\Phi} \approx \frac{D-26}{6} + \alpha^{'}(\nabla^2\Phi + \nabla_{\omega}\Phi\nabla^{\omega}\Phi) + O(\alpha^{'2}) = 0, \tag{23}$$

and

$$\beta^{G} \approx \alpha^{'}(R_{\mu\nu} + 2\nabla_{\mu}\nabla_{\nu}\Phi) + O(\alpha^{'2}) = 0. \tag{24}$$

The latter is a third example of how string theory entails the EFEs, in this case when gravity couples to a dilaton field. The simplest solution has a Minkowski target spacetime, $G_{\mu\nu}(X) = \eta_{\mu\nu}$, in which case

$$R_{\mu\nu} = 0 \quad \text{and} \quad \Phi(X) = V_{\mu}X^{\mu}, \tag{25}$$

where V_{μ} is a constant, so that the dilaton has a simple linear dependence on spacetime. But now (23) yields:

$$D = 26 - 6\alpha^{'}V_{\mu}V^{\mu}. \tag{26}$$

Earlier we saw that in the absence of a dilaton field, Weyl symmetry requires that $D = 26$ – the 'critical' dimension. Now we see that in the presence of a dilaton field D can take on other values, less than or greater than 26 (depending on whether the gradient of the dilaton is spacelike or timelike). Thus the dilaton relaxes the consequences of Weyl invariance, as we mentioned earlier. It is interesting to notice that the dimension now appears to be a dynamical (though constant) feature of the theory, controlled by the (square of) the gradient of the dilaton $V_{\mu}V^{\mu}$.

Now that we have introduced the linear dilaton, we need to indicate why controlling dilaton divergences requires Weyl invariance. Why? Recall, when we attempted to break Weyl symmetry the conformal factor ω became a linear dilaton field, $\Phi(X)$. Moreover, the dilaton is related to the coupling, $\alpha^{'} \sim e^{2\Phi(X)}$, so that its behavior signals the breakdown of perturbation theory. Thus violating Weyl symmetry requires that the dilaton divergences be controlled – but we shall now sketch how that itself requires Weyl invariance, so *the attempt to violate the symmetry fails.*

Consider a spacelike dilaton, $\Phi(X) = V_1X^1$. [Polchinski (2003), §9.9] notes that fixing divergencies at large X^1 can be achieved by introducing a 'tachyon profile', $\tau(x)$, into the action as a background field:

$$S_X - \frac{1}{2}\int d^DX\sqrt{-g}e^{-2\Phi}(g^{\mu\nu}\partial_{\mu}\tau_{\mu}(X)\partial_{\nu}\tau(X) - \frac{1}{\alpha^{'}}\tau^2(X)), \tag{27}$$

where S_X is the effective spacetime action (22). The equations of motion for the tachyon are then

$$-\partial_\mu \partial^\mu \tau(X) + 2V^\mu \partial_\mu \tau(X) - \frac{4}{\alpha'}\tau(X) = 0, \tag{28}$$

whose solution is

$$\tau(x) = exp(q \cdot X^1) \quad \text{where} \quad q = \sqrt{(\frac{D-26}{6\alpha'})}. \tag{29}$$

Once again the target space action can be rewritten as an equivalent worldsheet action, this time of the form

$$S_\sigma = \frac{1}{4\pi\alpha'} \int_\Sigma d^2\sigma \sqrt{-\gamma}[(\gamma^{ab}\eta_{\mu\nu}\partial_a X^\mu \partial_b X^\nu + \alpha' R V_1 X^1 + \tau_0 exp(q_1 \cdot x^1)], \tag{30}$$

where the dependence on D of the action (31) is contained in

$$q_1 = \sqrt{(\frac{26-D}{6\alpha'})} - \sqrt{(\frac{2-D}{6\alpha'})}. \tag{31}$$

For $D \leq 2$, q_1 is a positive quantity, so for $X^1 \to +\infty$ the tachyon exponential gets large, suppressing this limit in the path integral, yielding an effective repulsive potential. But it is precisely for $X^1 \to +\infty$ that the coupling

$$\alpha' \sim e^{2\Phi(X)} = e^{2\Phi(X) = V_\mu X^\mu}$$

diverges. Hence the tachyon controls the theory for $D \leq 2$. For $D > 2$, q_1 is complex, and the tachyon exponential is oscillatory, so the argument no longer holds. There's a dichotomy: either (as some have argued, [Polchinski (2003), 324]) some other mechanism is in play that prevents $X^1 \to +\infty$, or the theory breaks down. But either way, or if $D \leq 2$, the only hope for a linear dilaton – *and hence the only hope for the violation of Weyl symmetry* – is a tachyon field, (28). But, according to [Polchinski (2003), 323], (28) is the condition for Weyl invariance of the dilaton, and hence Weyl invariance is unavoidable (unless the divergences can be controlled in another way).

5. Conclusion

In this paper we have attempted to sketch enough of string theory to sketch the significance of conformal symmetry. Of course this is a huge and complex subject, and we can barely claim to have scratched the surface. However, we have indicated its crucial consequences, and also its necessity. Thus we hope to have established our principal claim, that conformal symmetry of string theory deserves to be a focus of attention in the philosophical

study of quantum gravity – itself one of the most pressing subjects within philosophy of physics.

References

[Becker et al. (2006)] Katrin Becker, Melanie. Becker, and John H Schwarz, *String theory and M-theory*. Cambridge University Press, 2006.

[Callan et al. (1985)] Curtis G Callan, D Friedan, EJ Martinec, and MJ Perry, "Strings in background fields", *Nuclear Physics B*, 262(4):593–609, 1985.

[Friedan (1980)] D. Friedan, "Nonlinear models in $2+\epsilon$ dimensions", *Physical Review Letters*, 45(13):1057–1060, 1980. doi: 10.1103/PhysRevLett.45.1057.

[Gasperini (2007)] Maurizio Gasperini, *Elements of string cosmology*. Cambridge University Press, 2007.

[Huggett and Wüthrich (2013)] Nick Huggett and Christian Wüthrich, "Emergent spacetime and empirical (in)coherence", *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 44(3):276–285, 2013.

[Kiritsis (2011)] Elias Kiritsis, *String theory in a nutshell*. Princeton University Press, 2011.

[Nakahara (2003)] Mikio Nakahara, *Geometry, topology and physics*. CRC Press, 2003.

[Polchinski (2003)] Joseph Gerard Polchinski, *String theory*. Cambridge University Press, 2003.

[Zwiebach (2004)] Barton Zwiebach, *A first course in string theory*. Cambridge University Press, 2004.

" Realism about the complexity of physical systems without realist commitments to their scientific representations: How to get the advantages of theft without honest toil"[1]

Cyrille Imbert
research fellow (CNRS, Archives Poincaré, Nancy), tenured.
email : cyrille.imbert@univ-lorraine.fr
institutional page: http://poincare.univ-lorraine.fr/fr/membre-titulaire/cyrille-imbert

Warning. While I try to defend in this paper a realist claim within general philosophy of science, the argument relies on basic but important notions in computer science. I did my best to write a self-content paper and present the argument and ideas in a non technical way so that the paper be accessible for philosophers with no training in computer science. In any case, all comments are welcome!

Version: March 2014. To be presented at PSA 2014, Chicago.

Abstract
This paper shows that, under certain reasonable conditions, if the investigation of the behavior of a physical system is difficult, no scientific change can make it significantly easier. This impossibility result implies that complexity is then a necessary feature of models which truly represent the target system and of all models which are rich enough to catch its behavior and therefore that it is an inevitable element of any possible science in which this behavior is accounted for. I finally argue that complexity can then be seen as representing an intrinsic feature of the system itself.

1. Introduction. The purpose of this paper is to show that, under certain reasonable conditions, if the investigation of the behavior of a physical system is difficult, no scientific change can make it significantly easier. This can be seen as some sort of impossibility result that says that some epistemic situation (in which investigating a system could be difficult for some agents and easy for others) cannot be met. It thereby shows that complexity is necessarily a feature of the model(s) (whatever it (they) turn(s) out to be) that truly represent(s) the target system, and of all models that are rich enough to catch its

[1] I am grateful to John Norton, Hervé Zwirn, Jacques Dubucs, Mikaël Cozic, Alexei Grinbaum and Mauricio Suarez for past or recent discussions about various versions of this argument. Remaining shortcomings are mine.

behavior. The complexity of the model can then be seen as representing an intrinsic characteristic of the system itself.

Though the idea supporting the claim is quite straightforward, it needs to be spelled out with great care since the validity of the argument hinges on its details. I first sketch this general idea in section 2 and present the thrust of the argument. Sections 3-5 are then devoted to the discussion of the steps and scope of the argument. I argue in particular that, in the described situations, complexity can be claimed to be an inevitable feature of any (mathematically possible) investigation of the corresponding systems.

2. Statement of the Problem and Sketch of the Argument. Progresses in mathematics or physics often make easier scientific tasks that were difficult, if not impossible, to carry out. For example, to decide whether a proposition of propositional logic is a tautology, one may laboriously enumerate all the 2^k possible cases; but once the tree method is known, things become much easier. Such progresses may originate both in mathematical findings and in advances in the empirical sciences such as the development of new modeling schemes. The invention of the boundary layer by Prandtl seems to illustrate this latter case. Navier-Stokes equations were derived in the mid-19th century but, because of their elliptic behavior, solving them was not possible for most practical problems like calculating the lift and drag on the first airplanes. By contrast, equations for the boundary layer were found to have parabolic behavior. This afforded significant analytical and computational simplification and the calculation of aerodynamic drag became possible for various situations. In such cases, the difficulty initially met by scientists is *epistemic* in the sense that it results from some lack of knowledge: Prandtl's invention showed that the investigation was *apparently* complex but not *intrinsically* so.

Is complexity always epistemic – and can we always entertain the hope that it may be swept away by scientific progress? Conversely, if complexity is not always epistemic, in which cases should it be seen as an objective feature of an investigation? People familiar with logic and computer science know that it can sometimes be proved that some mathematical tasks are genuinely difficult or impossible to carry out. The purpose of this paper is to make a riskier step by presenting cases in which *physical systems* can reliably be described as inherently complex.

Complexity is a property of mathematical models or problems[2]. Accordingly, establishing that some complexity properties intrinsically characterize some physical systems seems to require, at the least, showing that the corresponding models are *true* representations of these systems – and thereby solving the realism/anti-realism problem in such cases. A particularity (oddity?) of the argument to follow is that no such thing is in fact needed: the complexity of the models will be shown to characterize systems faithfully even if these models are false.

To make things clearer from the start, I shall now present a short version of the argument. Here are the assumptions. One wants to investigate the target behavior $B(S_i)$ of systems S_i, that is, of some set of systems of a common physical type S (e.g. Ising-like

[2] See section 4 for more about this point.

system) in different configurations (e.g. the numbers of spins, geometry and external fields may vary). A general model M (like "the" Ising model) yields the family of particular models M_i which is empirically adequate[3] regarding behavior $B(S_i)$. Finally, solving models M_i corresponds to a mathematical problem Π having irreducible computational complexity K. Let us now make the hypothesis that there exists another family of models M_i* which is also empirically adequate regarding behavior B_i and corresponds to a *simple* mathematical problem Π*.The claim is that this latter hypothesis implies a contradiction. Indeed, if Π* is simple, it should be possible, when trying to solve models M_i, to solve the corresponding easy models M_i* instead. It then becomes possible to solve problem Π easily, which, by assumption, is impossible. Therefore, if there is another empirically adequate family of models M_i* for $B(S_i)$, then the corresponding mathematical problem cannot be significantly simpler than Π.

Overall, the argument involves the main following claims:
(1) the investigation of the behavior of physical systems can be described as computational problems (in the computer science sense);
(2) such computational problems can be irreducibly complex;
(3) if a computational problem corresponding to the solution of a family of empirically adequate models is complex, then any other such family corresponds to an equally complex computational problem.

I provide in section 2 evidence for claims (1) and (2) and discuss claim (3), which is the potentially controversial core of the philosophical argument, in section 3.

2. Physical Investigations and Complex Computational Problems. In this section I argue that investigations about physical systems can sometimes be adequately described by means of irreducibly complex computational problems. A computational problem is an infinite collection of tasks of a common type such as "Given two numbers p and q, find the value of their sum $p+q$". The instances of the problem are the specific tasks that are actually being carried out, e.g. "1+1", "1+2", "2+1", etc. To theoretically study the behavior $B(S_i)$ of a system S_i, scientists need to investigate the corresponding property $P(M_i)$ of a model M_i standing for S_i. Then, they tackle the following task T_i "Based on (a suitable description of) M_i, find (a description of) $P(M_i)$". Since the physical parameters of S can indefinitely vary, the generic study of $B(S_i)$ corresponds to an infinite number of tasks T_i of a common type and therefore to a computational problem. For example, generic physical investigations like "given a Ising system composed of $p \times q \times r$ spins, calculate its equilibrium properties" or "given the description of a classical gas of n particles at time t_0, find its state at time t_0+t " correspond to computational problems.

The next step requires showing that some computational problems having a physical

[3] Empirical adequacy usually characterizes theories regarding all observable phenomena. I here use this notion for (family of) models regarding some specific behavior.

interpretation are intrinsically complex. Fortunately, computer scientists and physicists complete themselves this step by applying computational complexity theory (hereafter CCT) to physical problems. A problem is regarded as inherently difficult if any algorithm that solves its instances requires significant resources. CCT formalizes this intuition by robustly quantifying the resources needed to solve problems and by identifying a hierarchy of robust complexity classes (like NC, P, NP, EXP, etc.). For example, a decision problem is NP-complete if its solutions can be verified in polynomial time (it belongs to NP) and any problem in NP reduces to it in polynomial time; a problem is P-complete if it can be solved in polynomial time by a Turing machine (it belongs to P) and every problem in P reduces to it through an appropriate reduction.

A crucial notion in defining complete classes is that of reduction. A reduction is an algorithm transforming one problem into another one. For example, multiplication reduces to addition (2x3 = 2+2+2) and, if you know how to add, you know how to multiply. Reductions can also be used to show that the reduced problems are not more difficult than the reducing problem – provided that the cost of the reduction is negligible. Typically, to prove that a problem is NP-complete, polynomial reductions are used. Overall, if a problem is complete for a complexity class, unless this whole class collapses to some lower class (which, in the NP-complete case, is believed to be unlikely), no algorithm can be found to solve it significantly more quickly – whatever our scientific progresses.

It is a fact that some (major) physical problems have been proven complete for some complexity classes. I shall present two. The Ising-model has played for decades a central role in the development of modern statistical physics (Baxter, 1982). Whereas Onsager solved the two-dimensional case in 1944, its three-dimension version resisted investigations for decades till Baharona (1982) proved that evaluating its partition function is a NP-hard problem. More simple physical problems, like lattice gas models, can be complete for lower complexity classes. The investigation of lattice gases started in the 70ies as attempts to solve the Boltzmann equation for extremely simplified gazes of particles with discrete velocities (Hardy et al., 1973). Further inquiries proved that lattice gazes could be used to simulate Navier-Stokes equations (Frisch et al, 1986) and exhibit physical behavior. Lattice methods are currently being used in computational fluid dynamics for various applications such as the investigation of air flowing over vehicles. They have been proven P-complete (Moore and Nordhal, 1997), which essentially means that sequential polynomial simulations are required to investigate them.

We have reached so far the conclusion that, unless the complexity hierarchy partly collapses, the investigation of some physical models (e.g. Ising-like systems or billiard ball models) is irreducible complex (the degree of complexity being defined by the complexity class these models belong to). This conclusion calls for three remarks.

First, the conclusion reached so far is not that, *within some scientific practice,* some physical models are *actually* investigated by solving some complex computational problems – otherwise, one could answer that *these practices* are complex ones and involve difficult tasks like computing Fourier transforms, inverting matrixes, finding optima, etc. but that maybe one is using sledgehammers to crack nuts and the difficulty may be bypassed by finding simpler techniques. But the claim is that the mathematical problem –

versus some practices solving it – is a complex one. Intuitively, saying that a problem is inherently complex means that solving it requires large resources, *whatever the algorithm that is being used*, which already involves quantifying over possible methods. For example, if P≠NP, no algorithm can solve a NP-complete problem in polynomial time. Further, any problem of a low class of complexity can be solved via a complete problem of a high class of complexity, since low complexity classes are included in higher complexity classes. For example, deciding whether a number is even *can* be performed by reducing this problem to 3-SAT (a NP-complete problem) and solving instances of 3-SAT. But evenness is a simple problem and the complexity of 3-SAT does not lie in the set of instances that can be used to solve instances of evenness. By contrast, the results described above indicate that solving the Ising-model *is* NP-complete, which means that no polynomial algorithm can solve all its instances.

Second, when saying that some *physical* model has such or such complexity, I mean that all the instances of this model have a physical interpretation, which is the case for the Ising model and lattice gazes. If it were not so, the complexity of the computational problem may sometimes lie in a set of instances having no physical interpretation and then it would apparently but unduly characterize the physical problem.

It may however be rightly objected that using a complex model to study the behavior of a system does not imply that no simpler model can be used for the same purposes nor that complexity characterizes the system itself. Therefore, the argument still falls short of proving the claim that the complexity of the model faithfully represents some property of the system. Accordingly, to substantiate the realist claim, there is the need for an additional semantic assumption about the felicitousness of the representational relation between the models M_i and the target systems so that the features of the mathematical models be "tacked" to the physical systems. As we shall now see, the sweet aspect of the argument is that truth is by no means required to complete this step and empirical adequacy, a notion usually considered as innocuous and deceptive by realists, is sufficient to do the job.

3. The Core of the Argument. Let us now discuss the core of the argument. To put things briefly, it is assumed that an irreducibly model is used to study a system, that this model is empirically adequate and then it is shown by *a reductio ad absurdum* that it is not possible that another empirically adequate (possibly true!) *and* simple family of models does the same work.

Notations are as above. P_i (resp. P_i*) is the property of model M_i (resp. M_i*) that stands for behavior B_i and instances of problems Π (resp. Π*) are questions about these properties.

Assumptions regarding our epistemic situation:

- H1. Possibility of our practice. Target behavior B_i is in practice observable and models M_i can be in practice described (simplicity of modeling) and, when studying B_i, their content (once identified, see H3) can be meaningfully (simplicity of physical content description) ascribed to their target systems (simplicity of reference).

- H2. Semantic assumption. Family of models M_i is empirically adequate for behavior B_i.
- H3. Mathematical complexity assumption. Computational problem Π has irreducible complexity K.

Assumptions about the existence of another possible epistemic situation:

- H4. Semantic assumption. Family of models M_i* is empirically adequate for behavior B_i.
- H5. Mathematical complexity assumption. Computational problem Π* has complexity K* and K* is significantly lower than K in the CCT sense (e.g. Π* belongs to P and Π to NP).

H1 guarantees that we can easily ascribe the studied appearances to the target system and that the complex models we are using to investigate them are not *ad hoc* unduly intricate ways to investigate and refer to systems S_i. H2 says that these models are empirically successful: property P_i catches behavior B_i, that is, M_i has an empirical substructure that is isomorphic to appearance B_i even if the underlying theoretical description is false. H3 adds that these models correspond to an irreducible class of complexity (in the CCT sense). H4 says that another modeling practice is possible, which is not controversial, since any family of models isomorphic to models M_i will do. Strictly speaking, P_x and P_x* need not be the same properties since they catch target behavior B_i *up to isomorphism* (e.g. models M_i and M_i* may correspond to different reference frames). H4 and H5 jointly say that there exists another empirically adequate family of models that is *in addition* simpler to solve (in the CCT sense).

Here is now the *reductio ad absurdum.* Because of the empirical adequacy of the families of models M_i and M_j* for behavior B_i, questions about M_i (regarding P_i) have the same answers as those about the model M_j* (regarding P_j*) that represent the same systems – up to isomorphism. It is then tempting to use the instances of problem Π* to solve the instances of problem Π quickly. For any instance i of Π, one then needs to solve the associate instance j of Π*, that is, to solve model M_j* instead of model M_i. All it takes is to be able to identify for each M_i the corresponding M_j, or, in computational terms, to find a matching procedure that, given the description of M_i, translates it into the description of M_j* and thereby reduces problem Π to problem Π*. Since P_i and P_j* represent the same behavior B_i up to isomorphism, if P_j* is to be used to solve instances of Π, there may sometimes be the need to translate back the description of P_j* into the description of P_i.

Overall, the indirect way to solve models M_i (and problem Π) is composed of three steps, the matching procedure, the solution of models M_j* and, if necessary, the final return translation of the result. Here is now the catch. Since Π has irreducible complexity *K* and complexity cannot vanish in the air, one of these steps at least must also have complexity *K*. In brief, the original complexity constraint in the models that are actually being used has the following consequence:

(a) Either, other empirically adequate models M_j* have the same complexity *K*;
(b) Or their complexity is lower than *K* but matching models M_i with models M_j*

(if this is possible) has complexity *K* (e.g. if the original problem Π is NP-complete (resp. P-complete), then the translation procedure must be at least as costly, since Π* problem is comparatively easy).

(c) Or models M_j* and the matching procedure have complexity lower than *K* but translating back the description of P_j* into the description of P_i has complexity *K* – though the two statements say the same thing, up to isomorphism.

Since we are investigating the possibility of the existence of a simple *and* empirically adequate family of models M_j*, we need to analyze whether situations (b) or (c) are possible.

Let us first discuss case (b). The situation is the following. The two families of well-defined structures M_i and M_j* model the same family of systems and account for the same phenomena. Indeed, such correspondences between families of representations do exist in scientific practice, for example descriptions in Newtonian mechanics and Lagrangian mechanics, or descriptions in different reference frames, or standard representations and their Fourier transforms. The specificity of the situation is that one family of representation is (by assumption) intrinsically complex and the other simple. (Please note that when one makes a Fourier transform to make a calculation easier, one thereby proves that the original problem was not intrinsically difficult since you could transform it in a simpler problem). Finally, the investigated assumption is that matching the former to the latter is as difficult as solving the former, or even impossible: Is that latter assumption plausible?

Suppose that the simple family M_j* of models is *in practice usable for modeling purposes*, that is, that it is easily possible to match a non-theoretical characterization of system S_i to the corresponding model (simplicity of modeling assumption). Then, it seems that one can always find a matching procedure between the two families of models: starting from models M_i, come back to its pre-theoretical identifying description (simplicity of reference, see hypothesis H1) and then remodel the system within the modeling framework of models M_j*. Going through the shared pre-theoretical description is a way to establish some translation between the two types of description. Now suppose that it is possible to faithfully describe this translation procedure algorithmically. Then there is a contradiction because the procedure is algorithmic and simple whereas it was supposed to have irreducible complexity *K*.

The other option is to suppose that the translation procedure, *which can be cognitively carried out by modelers*, is some mental operation that is irreducibly not algorithmic in the way it is carried out (even if it de facto computes the matching between models M_i and models M_j*). Then, we are compelled to accept that some mental modeling operation can, by some magic, quickly solve (all possible instances of) a complex (P-complete, NP-complete, etc.) problem. As far as I know, there is no serious evidence in favor of this general possibility.

Overall, this means that if the complexity really lies in the matching procedure between families M_i and M_j, then there does exist simple models M_j but the modeling procedure to identify them must be as difficult as solving a problem with complexity *K* – for example solving a NP-complete problem if we are in the Ising case. Such a family of easy models then floats in the mathematical realm out of our modeling reach – and it can hardly

correspond to some possible-in-practice science.

It is worth insisting here: the modeling task does not lie *in the invention* of a new type of model. We can assume that models M_j are of a known type; what is here supposed to be difficult is the *standardized application* of this model type to physical situations of a known type, that is, finding particular versions of a general model that is known to correctly represent some type of situation. For example, the modeling task is not to invent the Ising model but to find the particular versions of the Ising model for particular systems of a common type (e.g. ferromagnetic systems having this or that geometry and number of atoms) of which we already know that the Ising-model is a good representation.

While the situation just described is implausible, I unfortunately have no clean, simple and final argument showing that it is logically, mathematically, or physically impossible. I even suspect that it is possible to cook up weird logical *ad hoc* constructs, possibly based on some costly transformations of the original problem Π, which make this situation possible. Typically, one may build into the modeling procedure the difficult steps of the solution of M_i and end up with some string of symbols computationally close to the solution of M_i; one may then claim that these strings are models M_j* and the trick is played. One may however doubt that the trick is acceptable, since all the complexity has been in practice transferred in the description of the family of models. Indeed, computer scientists do not seem to accept such descriptive procedures. Papadimitriou (a prominent computer scientist) notes: "There is a wide range of acceptable representations of integers, finite sets, graphs, and other such elementary objects. They may differ a lot in form and succinctness. However, all acceptable encodings are related polynomially. <…> In the course of this book, when we discuss a Turing machine that solves a particular computational problem, we shall always assume that a reasonably succinct input representation <…> is used" (1994, 26).

As philosophers of science, we may also add the acceptability constraint that the description of models M_j* should be made in a language that is suitable not only for investigating systems S_i but also other classes of systems – as can be expected from a language a) that is used within some general scientific theoretical practice which goes beyond the particular study of the systems the complexity of which is being discussed and b) that is appropriate to describe natural kind predicates.

Overall, and even in the absence of a formal proof, it seems safe to conclude that the matching procedure between family of models M_j and M_j, if these models are to be given acceptable descriptions, can hardly have complexity K – especially if we are discussing models that have supra-polynomial complexity, like the Ising model (cases involving polynomial complexity are in a sense more difficult to treat because "acceptable" encodings are usually polynomially related).

There now remains the possibility that the complexity might lie in the translation between the descriptions of P_i and P_j* (case c above) but a critical discussion can be made along the same lines as above. Indeed, since models M_i, M_j* and the appearance say the same thing*, up to isomorphism*, the translation between the description of P_j* and P_i can go through the description of appearances. Then we would have a case of an easy family of models having some acceptable description (since the complexity is no longer supposed to

lie in the modeling procedure) but determining what the solution P_j* means about the isomorphic appearances it represents would be a complex problem. For similar reasons as above, this possibility also appears implausible and unacceptable.

Let us wrap up. Situations (b) and (c) describe would-be situation in which there are two different ways of modeling, one tractable, the other not, but there is something like a computational gap between i) these two possible practices, either in terms of identifying the easy models or of translating their solutions; ii) between non theoretical descriptions of the target systems (and their appearances) and the description of the easy models (or of their solutions) – and of course, the more intrinsically complex the original problem, the larger this gap must be. The argument above shows that such situations are extremely implausible or "non acceptable". The converse conclusion is that in such cases, it is extremely plausible that any acceptable empirically adequate family of models (including the true models) have the same complexity *K* as the models we are actually using.

4. Discussion. I now want to clarify a few points about the content, validity and scope of the argument.

i) Strictly speaking, complexity characterizes computational problems (and models), so it cannot be directly and meaningfully ascribed to a physical system. The precise claim is that, in the discussed cases, *all satisfactory representations of a system* cannot but have this complexity property – which is a high-order property, since it describes a common feature of all possible algorithms that solve some models.

ii) It can however be claimed that the corresponding systems have been characterized intrinsically. Indeed, not only is the complexity property a feature of their true representation, it also characterizes all the representations that can be used to investigate the target behavior. This second stronger statement secures the intrinsicalness claim since it does not make it relative to any particular representation and shows that it is an essential feature of any possible investigation of the system (*versus* a somewhat accidental and neutral feature of the system or its representation). Indeed, if using the true representation of the system to investigate its behavior was difficult but the difficulty could be sidestepped by using proxy representations, complexity would be a true but shallow, without epistemic effect and somewhat contingent feature of the system. By contrast, the claim is that its nature is such that it is intrinsically difficult to investigate it, whatever the nature and "degree of truth" of the representation.

iii) The claim made is immune to progresses in computer or physics. Typically, if a system is said to be inherently complex, the advent of quantum computers will not make it theoretically easier – even if, for practical purposes, solving it may be much faster. Going from Marathons to Athens by car is quicker than running all the way but it does not make the distance shorter. In the same way, quantum computers may remove computational constraints for scientists but it will change neither the complexity hierarchy nor the interest for refining low complexity classes and seeing which models and problems belong to them.

iv) In the argument, I did not have to specify whether the would-be families of empirically adequate and simple models were to be derived from the same theory or result from some more substantial theoretical change. Therefore, the result describes the limits of

the progresses possibly generated both by findings in modeling *and* theoretical revolutions.

v) I did not have to root my realist claims about complexity in the supposed truth of some aspects of some representations. Thus, whereas most discussions of realist claims need to bring answers to anti-realist arguments (see for example Psillos, 1999), the present argument is noncommittal about but compatible with the validity of anti-realist arguments, like those in terms of pessimist induction, under-determination or skepticism about inference to the best explanation. Therefore, anti-realists may also have to bite the bullet and be realist about the complexity of, say, Ising-like systems. But conversely, as far as I can see, the argument is also noncommittal about existing realist arguments regarding scientific representations.

vi) Since irreducible complexity cannot vanish mysteriously, anyone willing to defeat the argument need to explain where the complexity of the original models has gone and why no translation between models doing the same work is possible; if this ever happen, we will definitely learn something valuable about possible sciences.

vii) I have however claimed that if such simple families do exist, they can hardly be part of an actual tractable scientific practice. Accordingly, even if the reader refuses to buy the realist claim, she may still have to buy the inevitabilist claim about what usable representations of such systems must be like in any possible-in-practice science.

5. Conclusion Fluid dynamics problems tackled by Prandtl were difficult but boundary layer models made them tractable. If the above argument is valid, no such progress is to be expected when the family of models that represent some system is both empirically successful and corresponds to an intrinsically intractable problem. In such cases, complexity is presumably an unavoidable property of all its acceptable representations and therefore faithfully reflects an intrinsic and essential property of the system.

References

Baxter, Rodney. J., 1989, *Exactly Solved Models in Statistical Mechanics*, Academic Press Inc, 1st edition 1982.

Baharona, Francisco, 1982, "On the computational complexity of Ising spin glass models", *Journal of Physics A: Mathematical and General*, 15 (10), 3241-3253.

Frisch, U., B. Hasslacher, and Y. Pomeau, 1986, "Lattice-Gas Automata for the Navier-Stokes Equation", *Physical Review Letters* 56, nº 14, 1505-1508.

Hardy, J., Y. Pomeau, and O. de Pazzis, 1973, "Time Evolution of a Two-Dimensional Classical Lattice System". *Physical Review Letters* 31, nº 5, 276-279.

Moore, Cristopher and Nordahl, Mats G. 1997, "Lattice Gas Prediction is P-complete", Santa Fe Institute Working Paper 97-04-034.

Papadimitriou, Christos, 1994, *Computational complexity*, Addison Wesley.

Psillos, Stathis, 1999, *Scientific Realism: How Science Tracks Truth*, London: Routledge.

To be presented at the PSA 2014, Chicago, 6-8 November 2014.
Please do not cite without permission.

On the Limits of Causal Modeling: Spatially-Structurally Complex Phenomena

Marie I. Kaiser

Word count: 7.170

Abstract

This paper examines the adequacy of causal graph theory as a tool for modeling biological phenomena and formalizing biological explanations. I point out that the causal graph approach reaches it limits when it comes to modeling biological phenomena that involve complex spatial and structural relations. Using a case study from molecular biology, DNA-binding and -recognition of proteins, I argue that causal graph models fail to adequately represent and explain causal phenomena in this field. The inadequacy of these models is due to their failure to include relevant spatial and structural information in a way that does not render the model non-explanatory, unmanageable, or inconsistent with basic assumptions of causal graph theory.

1. Introduction

In recent decades major advances have been made in formalizing causation and causal inference (Spirtes, Glymour, and Scheines 2000; Pearl 2000) and in using these formalisms to address traditional philosophical issues such as scientific discovery and the nature of scientific explanations (Woodward 2003, Hitchcock and Woodward 2003, Woodward and Hitchcock 2003). At the heart of these formal theories lie causal models that involve elements such as causal graphs, probability distributions, Bayesian nets, and structural equations which satisfy certain conditions, most prominently the Causal Markov Condition (more on this in Section 2). Causal models are appreciated because they allow for inferring causal relations from observed probabilistic correlations, for predicting the effects of manipulations and interventions, and because they can be used for representing and explaining causal relationships in very general, formal terms.

Proponents of the causal modeling approach usually emphasize and exemplify the wide scope of this approach. In recent years several authors have, for instance, shown that the causal modeling approach can also be applied to mechanistic explanations in biology and medicine (Casini et al. 2011; Gebharter and Kaiser 2014; Clarke, Leuridan, and Williamson forthcoming; Gebharter forthcoming). Also Woodward's interventionist theory of causation and causal explanation (2003) that makes extensive use of causal graphs is supposed to be applicable to a very wide range of causal relationships, including those in the biological sciences (2010, 2011, 2013).

In this paper I agree that causal modeling is a powerful approach to formally represent, explain, and discover causal relations. However, I also think that its scope

should not be overestimated and that it is important to recognize also the *limits* of the causal modeling approach. This paper explicates one of these limits: the explanation of spatially and structurally complex biological phenomena.[1] According to my line of criticism, formal causal models fail to offer adequate causal explanations of biological phenomena that essentially involve *complex spatial and chemical-structural relations*. This failure is due to the fact that causal graphs only provide causal difference-making information of the sort: A change in the value of *X* would under suitable conditions change the value (or probability distribution) of *Y*. The explanations of some biological phenomena, however, seem to be richer than this: these explanations do not only represent causal relations but also and prominently spatial relations and biochemical structures (such as the conformation and chemical structure of macromolecules, the spatial orientation and fitting of macromolecules to each other, and the complementarity of chemical structures). Based on the analysis of a case study from molecular biology, DNA recognition and binding by gene regulatory proteins, I show that the formal tools of causal graph theory are too impoverished to model biological processes that involve complex spatial-structural relations.

Interestingly, Woodward (2011) has basically conceded this point, but he does not see this as a limitation of his causal modeling or interventionist approach to scientific explanation. Taking a relaxed stance, he argues that complex spatiotemporal information

[1] Other limitations may be that causal models fail to account for the complex dynamics that biological phenomena such as biological clock mechanisms involve (Weber manuscript) and that causal models provide a confusing and ontologically inadequate view of the entities and activities involved in biological mechanisms (Gebharter and Kaiser 2014).

can just be added back to the backbone of causal difference-making information and can be used to "organize" (2011, 423) or "fine-tune" (2013, 55) causal difference-making information. This paper shows that things are not that easy. Some biological processes involve complex spatial and chemical-structural relations that are central to explaining these processes but that cannot be represented in causal graph models without rendering the model un-explanatory, unmanageable, or contradictory.

I proceed as follows. In Section 2 I briefly review the core notions and assumptions made in the causal modeling literature such as the notion of a causal graph, a probability distribution, and the Causal Markov Condition. In Section 3 I introduce the case study on which my analysis relies by explaining the three kinds of fit that are involved in DNA-protein recognition and binding (Section 3.1) and by specifying what exactly the phenomenon is that biologists seek to explain and which constraints on the adequacy of modeling this phenomenon follows from this (Section 3.2). In Section 4 I construct a causal graph model of DNA-protein recognition and then point out the shortcomings it has (Section 5). To elaborate my argumentation I discuss two possible objections that a causal modeler could raise: first, one might argue that the proposed causal graph model it not good enough, but that it is possible to construct an alternative model that *does* include the relevant spatial-structural information and *is* adequate (Section 5.1), second, one might object that I am not making an interesting or novel point as the explanation of DNA-protein recognition is non-causal and causal graph theory was not intended to model non-causal explanations (Section 5.2).

2. Causal Modeling

The most frequently used causal models can be grouped into two kinds: causal Bayesian networks and structural equation models (which are distinct but closely related, cf. Spirtes 2010). In this paper I focus on causal Bayesian networks (which can also be called causal graph models) and leave structural equations aside. Causal graph models combine mathematics and philosophy: the mathematical elements are directed acyclic graphs (DAGs) and probability theory (with focus on conditional independence); the philosophical elements are assumptions about the relationship between causation and probability (Spirtes, Glymour, and Scheines 2000).

A *directed acyclic graph* (DAG, also called *G*) is an ordered pair $G = \langle V, E \rangle$, where *V* is a set of variables and *E* is a set of directed edges (that are graphically represented by arrows) that have the variables in *V* as their vertices. A variable can be binary, its values representing for instance the instantiation or non-instantiation of some property, or variables can have multiple values or even be continuous. Here is an example of a DAG:

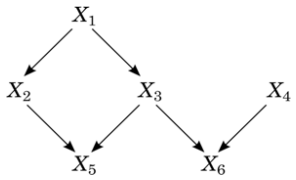


Fig. 1: An example of a directed acyclic graph (DAG)

Besides the DAG, a causal model consists of a *probability distribution P* over the variable set *V* that assigns a value to every variable in *V* such that the sum of all assigned variables in *V* equals 1. The pair of DAG *G* and probability distribution *P* over *V* is referred to as a

Bayesian network if and only if *G* and *P* satisfy the Markov Condition.[2] A DAG becomes a *causal graph* as soon as its edges are interpreted causally: an edge leading for instance from variable X_2 to variable X_5 (see Figure 1) is interpreted such that X_2 is a direct cause of X_5. When DAGs are interpreted causally the Markov condition (and d-separation) are assumed to be the correct connection between causal structure and probabilistic independence. This assumption is called the *Causal Markov Condition* and it can be formulated as follows:

Causal Markov Condition (CMC)

G and *P* satisfy the Causal Markov Condition if and only if for every variable *X* in *V*, $\text{INDEP}_{Pr}(X, \text{N-Des}(X) \mid \text{Pa}(X))$.

In words, a directed acyclic graph *G* and a probability distribution *P* over variable set *V* satisfy the Causal Markov Condition iff every variable *X* in *V* is probabilistically independent of *X*'s non-descendants (N-Des(*X*)) given *X*'s parents (Pa(*X*)). For example, X_2 is probabilistically independent from its non-descendants (i.e. X_3, X_4, and X_6) given X_1. CMC captures the intuition that conditioning on all common causes (e.g., X_1) and on intermediate causes breaks down the probabilistic influence between two formerly correlated variables (e.g., X_2 and X_3). Since causal models satisfy CMC they allow, for instance, for probabilistic prediction and manipulation.

[2] In DAGs the Markov Condition turns out to be equivalent to a more generally useful graphical relation: d-separation (Pearl 1988).

In the remaining sections of this paper I apply these formal tools of causal graph theory to a paradigmatic example of a biological explanation: the explanation of how gene regulatory proteins recognize and bind to a specific DNA region. My analysis will show that the causal modeling approach reaches its limits when it comes to representing and explaining spatially-structurally complex biological phenomena.[3]

3. DNA-Recognition and -Binding by Proteins

The regulation of the expression of genes is an important process in living beings. Differential gene expression is for instance the basis for cell differentiation during development. In eukaryotes, gene expression is regulated at different steps, for instance, a cell controls when and how often genes are transcribed (transcriptional control). The most important elements in regulating gene transcription are *gene regulatory proteins* (also called transcription factors). These proteins can recognize and bind to specific nucleotide sequences without having to open the double helix. This is due to the fact that the surface

[3] One might object that even in biological fields where spatially and structurally complex phenomena are studied (e.g., in protein folding and interaction research) graphical models turn out to be useful (cf. Balakrishnan et al. 2010, Thomas et al. 2009). I don't think, however, that cases like these are counterexamples to my thesis. First, the applied graphical models are *undirected probabilistic* graphical models, not direct causal graph models of the kind I discuss here. Second, in these studies graphical models are not used to directly *represent* or to *explain* phenomena such as protein folding or protein-protein interaction. Rather, they are used as techniques or tools, for instance, to guide the design of new protein sequences. Hence, graphical models might be useful in that domain, but not for the purposes I discuss in this paper.

of the DNA (in particular, its major grooves) presents a distinctive pattern of hydrogen bond donors and acceptors, and hydrophobic patches. A gene regulatory protein recognizes a specific DNA sequence because its surface is extensively complementary to the special surface features of the major groove of DNA. In other words, the protein *fits* well to a certain region of the DNA.

3.1 Three Kinds of Fit

This fitting of a gene regulatory protein to a specific DNA binding site can be interpreted as involving three interwoven aspects: a spatial aspect, a structural aspect, and a causal aspect. The *spatial fit* refers to the fact that the spatial conformation of the protein is such that it allows certain parts of the protein being placed in the major groove of the DNA double helix, at the DNA backbone, or in its minor groove. The *structural fit* means that the protein exhibits particular amino acid residues at certain places such that they are complementary to the functional groups that the nucleotide bases of the DNA exhibits at certain places. The spatial and the structural fit give rise to a *causal fit*. That is, the spatial orientation and the structural complementarity of the protein and the DNA enable that certain causal interactions between the function groups of the protein and those of the DNA take place and certain chemical bindings between them are established. These three kinds of fit can be summarized as follows:

(1) **Spatial fit**: The spatial conformation of the gene regulatory protein matches the double helix conformation of the DNA.

(2) **Structural fit**: The chemical structure of the protein (i.e., the sequence and location of its amino acids) is complementary to the nucleotide sequence of the DNA binding site.

(3) **Causal fit**: Certain amino acid residues causally interact with/make contact with certain nucleotide bases.

Consider the example of DNA recognition and binding by a *zinc finger* (ZnF). Zinc finger proteins are specific gene regulatory proteins that use zinc finger motifs to bind to DNA. The three zinc fingers of the Zif268 protein in mice, for instance, are arranged in a semicircular, C-shaped structure so that the α-helix of each zinc finger fits directly into the major groove of the DNA double strand (Pavletich and Pabo 1991). Figure 2 illustrates this spatial fit.

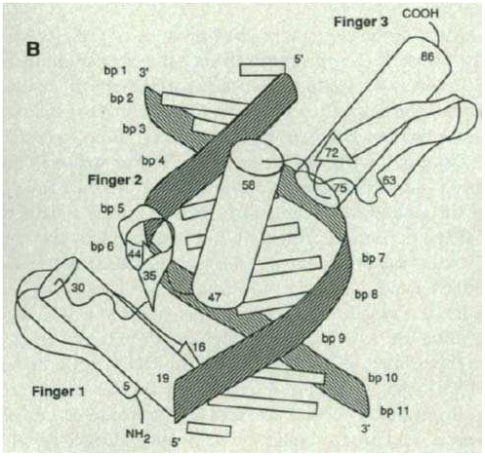


Fig. 2: Spatial fit between Zif268 and the DNA. (Pavletich/Pabo 1991, 811, Fig. 2.)

The cylinders and ribbons mark the α-helical and β-sheet regions of each finger. The shape of the Zif268 protein is such that it wraps round the DNA and the three α-helices (the cylinders) fit into the major groove.

The second kind of fit, the structural fit, is due to the fact that the chemical structure of the three zinc fingers is complementary to the nucleotide sequence of the DNA binding site. This means that the zinc finger protein possesses the "right" amino acids at the "right" places. For instance, the twenty forth amino acid of Zif268 protein is arginine, which has a positively charged residue. If the protein collides with the DNA binding site (in the right orientation) arginine is close to the nucleotide guanine, with which is can form hydrogen bonds (see Figure 3.b). Hence, the spatial and structural match between the zinc finger protein and its DNA binding region enables that the two also causally match: given that Zif268 collides with the DNA in the right orientation its three zinc fingers form extensive, characteristic contacts with the nucleotide bases (primarily along the guanine-rich DNA strand). Each finger has a similar relation to the DNA and makes its primary contacts in a three-base pair subsite. A summary of the critical base contacts is depicted in Figure 3a.

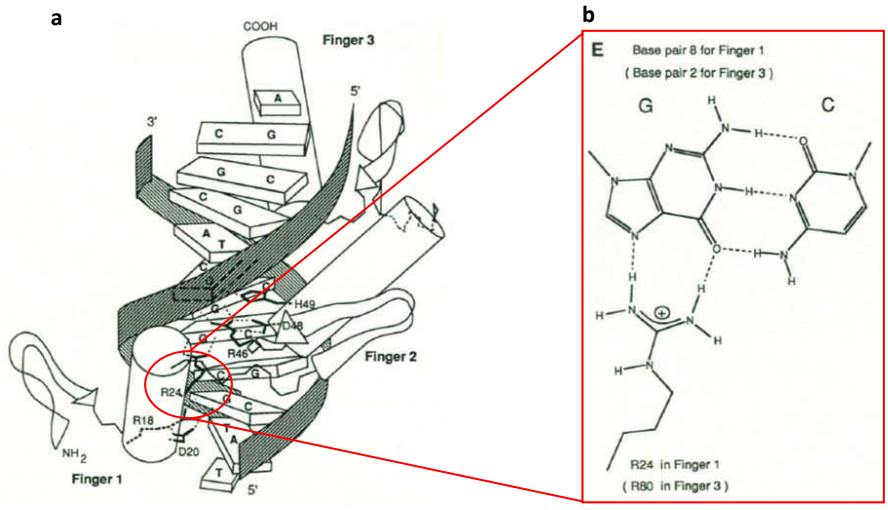


Fig. 3: Base contacts made by Zif268. (a: Summary of all critical base contacts; b: Arginine-guanine interaction that is present in finger 1 and 3; Pavletich and Pabo 1991, 812, Fig. 3.)

An example for such a contact between an amino acid residue of the zinc finger protein and a nucleotide base of the DNA is the Arginine-guanine contact (see Figure 3b). Arginine-guanine interactions seem to be responsible for much of the specificity in the zinc finger complex (Pavletich and Pabo 1991, 816).

3.2 Constraints on an Adequate Model of DNA-Protein Recognition

In this section I point out the implications that the biological literature has for the conditions under which an explanatory causal model of DNA-protein recognition is adequate. For this purpose I, at first, specify what exactly the phenomenon to be modeled

and to be explained in this case study is. Then I identify the different sorts of information that biologist treat as relevant to explaining this phenomenon. The underlying methodological assumption is that how biologists actually study and reason about DNA-protein recognition tells us which information is relevant to explaining this phenomenon and which information thus must be included in an adequate causal model.

In the biochemical literature the phenomenon is generally characterized as "DNA recognition by proteins" (Pavletich and Pabo 1991; Somers and Phillips 1992; Klemm et al. 1994) or as "protein-DNA interaction" (Luisi 1991). The target of these biochemical studies is to reveal *why* and *how* a particular gene regulatory protein (such as the zinc finger protein ZiF268 in mice) *recognizes* and *binds* to a specific DNA region.[4]

It seems to me that the phenomenon of DNA-protein recognition has (at least) two major characteristics, both of which must be captured by any adequate model of this phenomenon. First, a model of DNA-protein recognition must account for the *regular changes* from unbound proteins to DNA-bound proteins that take place under certain conditions. These regular changes include certain sub-processes, as the process of diffusion (of the protein and the DNA strand), the process of collision of the protein and the DNA, the process of recognition, and the process of binding of the protein to the DNA. Second, an adequate model of this phenomenon requires that one accounts for the *specificity* of the binding process. That is, a model must elucidate why a certain gene regulatory protein

[4] In mechanistic terms one could say that they seek to uncover the mechanism for DNA-protein recognition and binding.

recognizes and binds to a *specific* DNA region, rather than to a different region with a different nucleotide sequence.

The biochemical literature reveals further constraints on how DNA-protein recognition is adequately modeled. In Section 3.1 I have argued that biochemists provide three kinds of information when they explain how a certain gene regulatory protein recognizes and binds to a specific DNA region: first, they disclose the three-dimensional structure of the gene regulatory protein (which α helices and which β sheets it has and how they are located to each other) and its spatial orientation on the double helix when it is bound to the DNA. This amounts to showing that there is a *spatial fit* between the protein and its DNA binding region. Second, they reveal the chemical structure of the involved macromolecules and show that the chemical structure of the protein surface is *complementary* to the chemical features of the DNA sequence, that is, that there is a *structural fit* between protein and DNA binding site. Finally, biochemists point out which functional groups of the amino acid residues causally interact with which functional groups of the nucleotide bases and what the chemical nature of these interactions is (whether they are hydrogen bonds, van der Waals interactions, salt bridges, etc.). Modeling DNA-protein recognition typically involves providing a complete list of the contacts that are made between protein and DNA (e.g., Luisi et al. 1991, 502f; Pavletich and Pabo 1991, 812-814; Klemm et al. 1994, 23-25). This contact list specifies the *causal fit* between gene regulatory protein and DNA binding site.

All three kinds of information are necessary parts of an adequate causal model that provides an understanding and explanation of how a gene regulatory protein recognizes

and binds to a specific DNA region. Neglecting some of these relevant kinds of information renders the model inadequate. At least such a model would be incomplete in a way that is disastrous for its explanatory power. If one for instance ignores complex spatial and structural information and represents only the causal interactions between the functional groups of protein and DNA, the regular changes from unbound proteins to DNA-bound proteins and the specificity of the binding process will remain obscure.

4. How to Model DNA-Protein Recognition by Causal Graphs

How can we construct a formal causal model of the binding of a gene regulatory protein (e.g., Zif268) to a specific DNA binding site? The framework of causal graph theory (recall Section 2; for an introduction see, e.g., Spirtes et al. 2000) seems to be a promising tool. One possibility to model DNA-protein recognition by causal graphs (and probability distributions) is to conceive this phenomenon as a chain of causal events that leads to the binding of the gene regulatory protein Zif268 to the corresponding DNA binding site. The preceding causal events would then be the diffusion of Zif268 into the nucleus, which leads to (or allows for) the collision between Zif268 and the DNA at the corresponding binding site, which in turn causes the binding of Zif268 to the DNA. This chain of causal events can be represented by the following causal graph model (let us call it M_1):

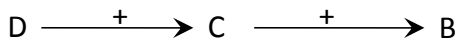


Fig. 4: Causal graph model M_1 for DNA-protein recognition.

D, *C*, and *B* are binary variables. *D* stands for the diffusion of Zif268 into the nucleus, *C* for the collision between Zif268 and the DNA at a certain region, and *B* for the binding of Zif268 to a certain DNA region. Each of the three variables can take one of the two values "taking place" and "not taking place". *D* is a direct cause of *C*, which is a direct cause of *B* (represented by the arrows). The "+" stands for a positive causal influence: raising the probability that *C* takes the value "taking place" raises the probability that *B* takes the value "taking place". Hence, the causal graph model allows for making predictions and testing them by interventions. For instance, the causal dependency relation between *C* and *B* can be tested by lowering the probability of the collision of proteins and DNA strands by dilution. According to the causal model the consequence of this intervention should be that also the probability for the binding of the protein to the DNA decreases.

A possible objection to M_1 is that it does not account for the causal fit between the gene regulatory protein Zif268 and the DNA binding site (not to mention the spatial and the structural fit). In the model there is only a single variable *B* that stands for the binding of Zif268 to a certain DNA region. No further information about *which* causal interactions between *which* amino acid residues and nucleotide bases take place and bring about the binding is included in the model. But this information seems to be crucial for an understanding of the causal fit between Zif268 and the DNA binding site. This objection can be avoided by choosing a larger number of more fine-grained variables. Instead of representing the collision of the entire protein with the entire DNA binding site by a single variable (*C*) this event is broken down to into various sub-events (e.g., amino acid residue x_1 collides with nucleotide base y_1, amino acid residue x_2 collides with nucleotide base y_2,

etc.), each of which is then represented by a distinct variable (C_1, C_2, etc.). Likewise, not the binding of the protein to the binding site in general is represented (by a single variable B). Rather, different variables (B_1, B_2, etc.) represent the binding of different amino acid residues to different nucleotide bases.[5] The resulting causal graph model M_2 is the following:

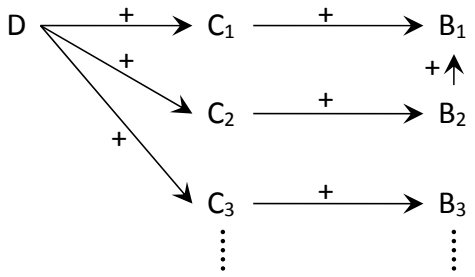


Fig. 5: Causal graph model M_2 for DNA-protein recognition.

C_1, …, C_n and B_1, …, B_n are binary variables. Each of them can take one of the two values "taking place" and "not taking place". C_1, …, C_n stand for the collision (or spatial proximity) of a certain amino acid residue of Zif268 with a certain nucleotide base, and B_1, …, B_n for the formation of a certain set of bonds/interactions between residues and bases.

[5] One might claim that these variables (C_1, C_2, C_3, B_1, B_2, and B_3) are not only more fine-grained, but also lower-level variables (i.e., variables that represent entities that are located on a lower organizational level than the entities represented by the former variables C and B). An even more "fine-grained" model would distinguish also among the different contacts that are made between one amino acid and one nucleotide.

The arrows between the variables represent that *D* is a direct cause of C_1, C_2, and C_3, which are direct causes of B_1, B_2, and B_3. Since the binding of a particular amino acid residue to a particular nucleotide base may raise the probability that another binding between the protein and the DNA is established the causal graph model also contains arrows between the *B*-variables (such as the arrow between B_1 and B_2).[6]

5. Where the Limits of Causal Graph Models Lie

In Section 3.2 I have argued that three kinds of information are crucial to adequately explaining why and how gene regulatory proteins, such as Zif268, specifically and regularly recognize and bind to certain DNA regions: information about the spatial fit, about the structural fit, and about the causal fit between protein and DNA. M_2 is superior to M_1 since it succeeds in representing the causal fit, that is, M_2 includes information about which amino acid residues causally interact with and establish chemical bindings to which nucleotide bases. However, M_2 fails to provide an understanding of the spatial fit as well as of the structural fit between Zif268 and DNA. It entails no information about the conformation of Zif268 and about how it matches the double helix shape of the DNA (spatial fit). It also does not represent the complementarity of the chemical structure of the protein (i.e., the sequence and location of its amino acids) to the nucleotide sequence of the DNA binding site (structural fit). M_2 includes information about which amino acids causally interact with which nucleotides (contained in the variables $B_1,\ldots B_n$). But this is, as such, no direct or complete information about the structural fit between protein and DNA:

[6] The same might be true regarding the *C*-variables.

The fact that protein and DNA causally fit allows only *inferences* about the complementarity of their chemical structures and it sheds light on only *some* parts of the chemical structure of DNA and protein. The causal graph model M_2 is thus inadequate because it leaves out spatial and structural information that is crucial for explaining why and how the zinc finger protein specifically and regularly recognizes and binds to DNA.

In the next two subsections I elaborate on my argument by addressing two possible objections that the causal modeler could raise. First, one might admit that M_2 is inadequate, but argue that it is possible to construct an *alternative causal graph model* that accounts for the spatial and structural fit between Zif268 and DNA and thus is adequate (Section 5.1). Second, one might agree that causal graph theory is not the appropriate tool to model spatially and structurally complex phenomena, but object that this is neither an interesting nor innovative insight as the explanations in these cases are *non-causal* and causal graph theory was not intended to model non-causal explanations (Section 5.2).

5.1 Shortcomings of Alternative Modeling Strategies

The first objection says that even if the causal graph model M_2 is inadequate it is possible to revise M_2 in a way that renders it adequate. In other words, an opponent might argue that it is possible to construct an alternative causal graph model that includes all the complex spatial and structural information that is relevant to explaining DNA-protein recognition into M_2. But how could that be done? I see two different ways one could go.

One option is to include information about the spatial conformation of Zif268 into the characterization of the variables $C_1, \ldots, C_n$ or $B_1, \ldots, B_n$. For instance, one could

characterize B_2 not as the "formation of a certain set of contacts between amino acid residue x_2 and nucleotide base y_2" but as the "formation of a certain set of contacts between amino acid residue x_2 and nucleotide base y_2, where x_2 is covalently bound to x_1 and x_3, forms a salt bridge to x_{14}, has a close distance also to bases y_3 and y_9, and so on".

But this strategy encounters several problems. First, storing complex spatial and structural information into the characterization of the variables renders M_2 *unmanageable* because the measurement of the values of the variables becomes very complicated or even unfeasible.[7] Second, this strategy results in a causal model in which the spatial and structural information is *highly fragmented* because information about the spatial conformation of Zif268 and of the DNA binding site, about their spatial orientation to each other, about the chemical structures of protein and DNA, and about the complementarity of these structures is distributed over the many variables $C_1, \ldots, C_n$ and $B_1, \ldots, B_n$. This fragmentation is devastating for the explanatory power of the causal model since the resulting model fails to elucidate why and how Zif268 spatially and structurally fits to the DNA bind site (e.g. that the whole protein Zif268 has a C-shaped structure so that the α-helix of each zinc finger fits directly into the major groove of the DNA double strand). Third, this strategy of storing complex spatial and structural information into M_2 gives rise to a causal model in which a great deal of the explanatorily relevant information is contained in the characterization of the variables, not in the represented causal dependency relations. This suggests that the causal dependency relations are less informative and bear

[7] One might even argue that spatial and structural information is so complex that adding them to the causal model is *not feasible in practice*, but merely possible in principle.

less explanatory weight than the characterization of the variables. But this seems to *conflict* with the typical way of how causal graph models are conceived and reinforces the impression that one tries to add something that does not really fit.

A second option is to add one or more variables to the causal model that are supposed to represent the spatial and structural fit between protein and DNA. One could, for instance, add the variable *S* that stands for the protein Zif268 having a certain conformation and chemical structure, the DNA having a certain shape and chemical structure, Zif268 being oriented towards the DNA binding site in a certain way, and the structures of protein and DNA being complementary.[8] In its simplest form *S* would be a binary variable, which can take one of the two values "being realized" and "not being realized". The resulting causal model M_3 would look as follows:

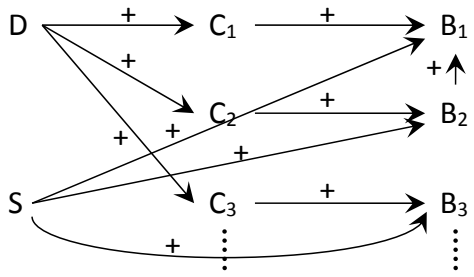


Fig. 6: Causal graph model M_3 for DNA-protein recognition.

[8] This option implies that one accepts that the conformations of, spatial relations between, and chemical structures of protein and DNA are *difference makers* with respect to their binding and that – given an interventionist, counterfactual theory of causation – information about the spatial and structural fit also is *causal* information (see Section 5.2).

M_3 accounts for the fact that whether or not a certain amino acid base makes contact to a certain nucleotide base depends not only on whether protein and DNA collide, but also on whether they spatially and structurally fit (i.e. on whether the amino acid base and the nucleotide base are complementary and whether they are proximate enough).

The revised causal model M_3, however, is still not satisfactory. First, it has the feel of *putting* the missing relevant spatial and structural information *in "by hand"*:[9] something that does not smoothly, automatically fit must be added under additional, atypical efforts. What reinforces this feeling is that all different elements of the spatial and of the structural fit are represented by a single variable *S*. It would be more adequate to add different variables (S_1, …, S_n) each for protein conformation, DNA shape, spatial orientation of protein towards DNA, amino acid sequence of Zif268, chemical structure of the DNA binding site, and for the complementarity of protein and DNA structure. These different variables could then also be quantitative variables (rather than binary ones), which might represent, for instance, relative distances among sets of protein molecules and sets of DNA molecules and possible combinations of proteins with a certain sequence binding to DNA molecules with a certain sequence. But this strategy encounters the problem that the variables are no longer *conceptually independent* from each other (e.g., a specific protein conformation requires a specific amino acid sequence, or the complementarity of protein and DNA binding site means that certain kinds of causal interactions can take place), which violates a central requirement of causal graph theory. Finally, even if we could add different variables for each element of spatial and structural fit the characterization of the

[9] Thanks to an anonymous referee for pointing this out.

variables would be very complex. For instance, the variable that stands for the complementarity of protein and DNA structure would have to be characterized in a way that shows not only *that* the protein structure is complementary to the structure of the DNA binding site, but that also illuminates *why* they are complementary and *what* this amounts to. This requires storing detailed information about which functional groups of which amino acid residues fit (for which reasons) to which functional groups of which nucleotide bases into the characterization of variables. The second option of including complex spatial and structural information into the causal graph model thus encounters the same objections as the first option: storing complex information into the characterization of the variables renders the causal model *unmanageable* and it *contradicts* the basic assumption of causal graph theory that causal dependency relations between variables are a *central* element of causal models (i.e. that they are not less informative and bear less explanatory weight than the variables themselves).

To conclude, even if it might be in principle or technically possible to include information about complex spatial relations and chemical structures into a causal graph model, this can only be done at the expense of the adequacy of the causal model as it renders the causal model non-explanatory (because the information gets highly fragmented), unmanageable, and entails inconsistencies with basic assumptions of causal graph theory (e.g. that variables must be conceptually independent and that causal

dependency relations are central). Hence, the causal graph approach reaches its limits when it comes to explaining spatially and structurally complex biological phenomena.[10]

5.2 Just a Matter of Non-causal Explanations?

A second line of criticism agrees with me that biological phenomena that involve complex spatial relations and chemical structures cannot be adequately modeled by the tools of causal graph theory, but argues that this is neither an interesting nor novel insight. Everybody agrees, so the criticism goes, that there are non-causal explanations in science such as the explanation of DNA-protein recognition and that *causal* graph theory is not the appropriate tool for representing such *non-causal* explanations. So what is the big news?

I agree that if the explanation of DNA-protein recognition were non-causal it would be weird to investigate why *causal* graph theory fails to adequately represent this *non-causal* explanation. But I think the process of DNA-protein recognition clearly is a *causal* process, which involves causal interactions between the functional groups of the protein and of the DNA binding site, and that this process must be explained causally, too.[11] What is interesting about this explanation is that, besides causal relations, it also and prominently

[10] Jantzen and Danks (2008) have recently argued that topological properties of complex molecules can be represented by graphical models. One might suggest that this challenges the argument that I provide in this paper. Note, however, that the graph models I discuss are very different from the ones that Jantzen and Danks propose. They use *undirected* graphs to represent topological *properties* of molecules, whereas I use *directed*, causal graphs to represent *processes* of DNA-protein binding.

[11] I regard explanations as causal if they explicitly (but not necessarily only) represent causal relations.

represents relations that are *non-causal* (or at least not directly causal): the shape of the gene regulatory protein Zif268, how Zif268 is spatially oriented to the DNA double helix (i.e. their spatial fit), the chemical structures of Zif268 and DNA binding site (i.e. the sequence of amino acids and nucleotides), and the complementarity of their chemical structures. The question of whether these kinds of non-causal information can be adequately represented in causal graph model or whether they constitute a limit of the applicability of the causal modeling approach is neither uninteresting nor has it already been sufficiently addressed.[12]

One might question whether information about the spatial and structural fit between DNA and protein is in fact non-causal as both kinds of fit *make a difference* to the binding of DNA and protein (in other words, the binding counterfactually depends on there being a spatial and structural fit). Assuming a counterfactual view of causation one could then argue that (besides the collision) the spatial and structural fit between DNA and protein *cause* their binding. It is important to note that my analysis of the limits of causal graph theory is *compatible* with such an interpretation. Characterizing complex spatial and structural information as causal makes it even more urgent and interesting to analyze whether these kinds of information can be included in causal models.

The discussion about the allegedly non-causal character of the explanation of DNA-protein recognition points to another important issue: in explanations of spatially and

[12] This question has already been discussed with regard to constitutive or part-whole relations (e.g., Casini et al. 2013; Gebharter and Kaiser 2014). This paper extends the discussion to spatial and structural relations.

structurally complex biological phenomena causal and non-causal information often cannot be easily separated, but rather are *interwoven* and highly integrated. For example, in my case study the structural fit between a gene regulatory protein and the DNA binding site is specified by referring to non-causal information such as information about the sequence of amino acids and nucleotides and to information about the complementarity of protein and DNA structures, which seem to be implicitly causal or to be closely connected to causal information. The complementarity of protein and DNA can be spelled out either dispositionally or counterfactually: to say that protein and DNA are complementary means to say that protein and DNA have the disposition to causally interact/form certain kinds of bounds (if they collide or are proximate enough) or that if protein and DNA collided or were proximate enough they would form certain kinds of bounds. Both explications refer to causal information about the formation of bounds between protein and DNA (i.e. information about the causal fit). Furthermore, the causal part of the explanation of DNA-protein recognition seems to rely on its non-causal part because to understand the causal fit between protein and DNA binding site *requires* understanding how they spatially fit together and that their structures match. Hence, even if it is possible to entangle three respects how a gene regulatory protein fits to a specific DNA binding site (recall Section 3.1) in fact these three kinds of fit and the causal and non-causal information they invoke are interwoven.

This entanglement of causal (difference-making) information with non-causal (spatial-structural) information poses a challenge to Woodward's argument that spatial information can simply be added to the backbone of causal difference-making information

and can be used to fine-tune causal information (2011, 2013). His argument requires that causal difference-making information and spatial-structural information are easily separable. But biological practice shows that in some cases information about causal and non-causal relations is closely related and interwoven in such complex ways that it is not possible to clearly tell them apart.

6. Conclusions

I agree that causal modeling is central to scientific practice and that formal theories of causal modeling and explanation, such as causal graph theory, are powerful. However, I think that their significance is not universal and that it is important to notice also the *limits* of causal graph theory. In this paper I have used an example from molecular biology to reveal one of these limitations: spatially-structurally complex phenomena. I have shown that causal graph models fail to provide explanations of biological processes that involve complex *spatial-structural* relations (such as DNA-protein recognition).

The goal of this paper has not been to argue that there exist kinds of explanation in the biological sciences that are non-causal. The process of how a gene regulatory protein regularly recognizes and binds to a specific DNA region clearly is a *causal* process. But these causal relations are not the only kind of relations that is relevant to understanding and explaining the phenomenon of DNA-protein recognition. Besides causal difference making information, any adequate causal model of this phenomenon must also and prominently include *spatial and structural information* (i.e. information about protein and DNA conformation and chemical structure, about how they spatially fit, about why their

structures are complementary). This non-causal (spatial and structural) information is often entangled with causal difference making information. But exactly this is, as I have argued, the point at which the *limits* of causal graph models become apparent.

My central argument in this paper has been that the formal tools of causal graph theory are inappropriate to model and to explain spatially and structurally complex biological phenomena (e.g. DNA-protein recognition) because they result in causal models which either ignore the importance of spatial and structural relations all together (such as M_1 and M_2) or which try to include the relevant spatial and structural information but, in so doing, render the causal graph models non-explanatory, unmanageable, or inadequate because they conflict with basic assumptions of causal graph theory (such as M_3). This argument does not erode the significance of formal approaches to causal modeling, but it demonstrates that their scope is limited.

References

Balakrishnan, S., Kamisetty, K., Carbonell, J. G., Lee, S.-I., Langmead, C. J. 2010. "Learning generative models for protein fold families." Proteins 79(4): 1061-1078.

Casini, L., P. M. Illary, F. Russo, and J. Williamson 2011. "Models for Prediction, Explanation and Control: Recursive Bayesian Networks." Theoria. An International Journal for Theory, History and Foundations of Science 70(1): 5-33.

Gebharter, A. and M. I. Kaiser 2014. "Causal Graphs and Biological Mechanisms." In Explanation in the Special Sciences. The Case of Biology and History by M. I. Kaiser, O. Scholz, D. Plenge and A. Hüttemann. Berlin: Springer, 55-85.

Gebharter, A. forthcoming. "A formal framework for representing mechanisms?" Philosophy of Science, 81(1).

Jantzen, B., Danks, D. 2008. "Biological Codes and Topological Causation." Philosophy of Science 75(3): 259-277.

Hitchcock, C. and J. Woodward 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." Noûs 37(2): 181-199.

Klemm, Juli D., Mark A. Rould, Rajeev Aurora, Winship Herr, and Carl O. Pabo 1994. "Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules". Cell 77: 21-32.

Luisi, B.F., W. X. Xu, Z. Otwinowski, L. P. Freedman, K. R. Yamamoto, P. B. Sigler 1991. "Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA". Nature 352:497-505.

Pavletich, Nikola P. and Carl O. Pabo 1991."Zinc Finger-DNA Recognition: Crystal Structure of a Zif268-DNA Complex at 2.1 A". Science, New Series 252:809-817.

Pearl, J. 1988. Probabilistic reasoning in intelligent systems. San Mateo CA.: Morgan and Kaufman.

Pearl, J. 2000. Causality. Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Somers, W. S. and S. E. Phillips 1992. "Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands". Nature 359:387-393.

Spirtes, P. 2010. "Introduction to Causal Inference." Journal of Machine Learning Research 11: 1643-1662.

Spirtes, P., Glymour, C. and R. Scheines. 2000. Causation, Prediction, and Search. Cambridge, Mass.: MIT Press.

Thomas, J., Ramakrishnan, N., Bailey-Kellogg, C. 2009. "Graphical models of protein-protein interaction specificity from correlated mutations and interaction data." Proteins 76(4): 911-929.

Woodward, J. 2003. Making Things Happen: A Theory of Causal Explanation. New York: Oxford University Press.

Woodward, J. 2010. "Causation in biology: stability, specificity, and the choice of levels of explanation." Biology & Philosophy 25: 287-318.

Woodward, J. 2011. "Mechanisms Revisited." Synthese 183(3): 409-427.

Woodward, J. 2013. "Mechanistic Explanation: Its Scope and Limits." Proceedings of the Aristotelian Society Supplementary Volume 87(1): 39-65.

Woodward, J. and C. Hitchcock 2003. "Explanatory Generalizations, Part I: A Counterfactual Account." Noûs 37(1): 1-24.

Weber, Marcel (manuscript): "On the Incompatibility of Dynamical Biological Mechanisms and Causal Graph Theory", presented at the PSA 2014, Chicago, 6-8 November 2014.

Unification and the Quantum Hypothesis in 1900–1913

Abstract

In this paper, I consider some of the first appearances of a hypothesis of quantized energy between the years 1900 and 1913 and provide an analysis of the nature of the unificatory power of this hypothesis in a Bayesian framework. I argue that the best way to understand the unification here is in terms of informational relevance: on the assumption of the quantum hypothesis, phenomena that were previously thought to be unrelated turned out to yield information about one another based on agreeing measurements of the numerical value of Planck's constant.

Word Count: 4,977

1 Introduction

The idea that unification is a virtue of a scientific theory has a long history in philosophy of science, and has been presented in several guises. Accounts range over those focused on the common causal origins of various phenomena to those emphasizing a common explanatory basis. Of course, these are not mutually exclusive ideas and a combination of these elements

is common. Examples include William Whewell on the Consilience of Inductions (1989), Michel Janssen on Common Origin Inferences (2002), Philip Kitcher on explanatory unification (1989) and William Wimsatt on robustness (1981). While the details of such discussions differ, it is clear that some version of this notion has played a role in several important episodes of scientific theorizing. One period in the history of science that has perhaps been neglected in this context is the case of the years leading up to the development of the old quantum theory. While the history is well-documented (e.g. Klein (1961, 1965, 1966), ter Haar (1967), Hund (1974), Kuhn (1978), Mehra & Rechenberg (1982)) and there have been excellent discussions of the justification of particular aspects of the theory (Norton (1987, 1993)), an explicit discussion in terms of unification has not yet been provided.

First, consider what I refer to as the quantum hypothesis, *QH*: this is the idea that radiation energy cannot always be treated in a continuous manner as in classical physics, but that instead, radiation of frequency ν is emitted and absorbed in packets of size $h\nu$, where h refers to the universal physical constant referred to as Planck's constant. I will consider some of the central applications of *QH* between 1900 and 1913 and explain its unificatory role. I will argue that the best way to understand the unification present in this period is in terms of informational relevance: on the assumption of *QH*, observations performed on various diverse physical phenomena can be thought of as measuring or constraining the numerical value of h, and these agreeing measurements of h render the physical phenomena relevant to one another by providing information about the value one is likely to obtain in the various cases.

Such a feature is particularly important in early stages of scientific theory development before alternative methods of justification are available. Despite the absence of a coherent theory that incorporated *QH*, and despite its inconsistency with well-established physics, it was taken by several scientists to be a promising starting point for the development of a more adequate

theory. In this paper I will argue that they had reasonable grounds for doing so, because of empirical support for *QH* in the form of its unificatory power.

I will first give an outline of unification in terms of informational relevance and comment briefly on how I take this concept to differ from a common cause argument. I will then give an overview of several scientific episodes invoking Planck's constant and in doing so, I will argue for my claim that the type of unification displayed here is best understood in terms of experiments on phenomena providing information about one another by yielding agreeing measurements of the parameter h.

2 Unification

As mentioned previously, there are several different ways to conceive of unification as well as differences in what is being attributed unificatory power. I defend the idea that the unification should be understood in terms of informational relevance. My claim is not that this necessarily captures scientists' actual motivations, but that we can retrospectively identify the fact that the quantum hypothesis had unificatory power in the way to be explained. Here, I have adopted the explication of unification given by Myrvold (2003).

Myrvold provides a Bayesian account of the feature of unification. He shows that on a particular understanding of what it means for a hypothesis to *unify* phenomena, its ability to do so contributes directly to its support by the evidence it unifies. Thus, if one accepts a Bayesian confirmational framework, the unifying hypothesis obtains support from the unifying phenomena.

More specifically, he takes a common definition of the *informational relevance* of a proposition p_1 to another proposition p_2, conditional on background b,

$$I(p_2, p_1|b) = Log_2 \frac{Pr(p_2|p_1\&b)}{Pr(p_2|b)}. \quad (1)$$

He then defines the quantity U as a measure of the extent to which h unifies p_1 and p_2,

$$U(p_1, p_2; h|b) = I(p_1, p_2|h\&b) - I(p_1, p_2|b). \quad (2)$$

This generalizes straightforwardly to a set of hypotheses $p_1 \ldots p_n$ (2003, 411). He then shows that on two common candidates for the degree to which evidence supports a hypothesis, the quantity U contributes directly to the support of h by the evidence. I use the "degree of confirmation" in my discussion, but a similar result holds if one takes Good's measure of the "weight of evidence." Notice that the degree of confirmation, measured by $log\frac{Pr(h|e\&b)}{Pr(h|b)}$, is identical to the definition of the informational relevance of e to h, so we can consider the informational relevance as a measure of evidential support. We thus obtain for the informational relevance of evidence e_1 and e_2 to hypothesis h,

$$I(h, e_1\&e_2|b) = I(h, e_1|b) + I(h, e_2|b) + U(e_1, e_2; h|b). \quad (3)$$

Myrvold explains the significance as follows:

> [T]he degree of support provided to h by e_1 and e_2 taken together is the sum of three terms: the degree of support of h by e_1 alone, the degree of support h by e_2 alone, and an additional term which is simply the degree of unification of the set $\{e_1, e_2\}$ by h. An analogous result holds for larger bodies of evidence. (2003, 412)

Thus, the ability of a hypothesis to unify previously unrelated phenomena contributes directly

to the likelihood of that hypothesis given the evidence. In what follows, I will provide details of how the case under consideration provides an example of this feature. Briefly, the physical phenomena to be discussed were not clearly relevant to one another before the postulation of *QH*. However, on the assumption of such a hypothesis, numerical values of quantities obtained from observations of those phenomena could be used to calculate the numerical value of h, all of which agreed to within an order of magnitude. The measured value of h via one type of phenomenon thus provided information about the measured value of h via a different phenomenon when assuming *QH*.

The explication of unification in terms of informational relevance certainly does not preclude the existence of a common cause argument, such as the one given by Wesley Salmon in his reconstruction of Jean Perrin's determination of Avogadro's constant (1984). However, I would argue that characterizing the unificatory power of *QH* as being due to a common cause would be to overstate the strength of the information available, since at the time in question, it was not at all clear how quantization might be occurring, and no account of the underlying mechanisms was forthcoming. Despite this, it was clear that *QH* did possess unificatory power in the informational relevance sense, and this minimal sense is all that is required to provide the hypothesis with some confirmational force.

3 Uses of the Quantum Hypothesis

3.1 Blackbody Radiation

Planck's constant, h, was first introduced by Planck in his work on blackbody radiation (1900). He began by interpolating an expression for the equilibrium entropy based on existing laws which were only partially empirically adequate; with this formula, he was able to

determine a radiation formula that correctly described the entire emission spectrum of blackbody radiation, $E = \frac{c_1\lambda^{-5}}{e^{c_2/\lambda T} - 1}$. The interpretation of this formula led him to posit the quantum hypothesis, that energy of frequency ν is absorbed and emitted in packets of size $h\nu$. The formula for energy distribution could then be written as

$$E = \frac{8\pi ch}{\lambda^5}\frac{1}{e^{ch/k\lambda T} - 1}. \tag{4}$$

Planck was then able to use this empirically confirmed radiation formula to estimate the size of h. He used his formula to calculate the amount of radiation in air, and compared this with values obtained by Ferdinand Kurlbaum in experimental work (1898). He then drew on observations made by Otto Lummer and Ernst Pringsheim, who were able to determine the wavelength of the maximum energy in air of blackbody radiation. The result was a numerical value for the parameter h, $h = 6.55 \cdot 10^{-27} erg \cdot sec$.

Near the beginning of his work on blackbody radiation, Planck was focused on providing observationally motivated descriptions of phenomena using a general idea of 'resonators' while hoping that electron theory would later be able to fill in the gaps, so to speak, on how absorption and emission of discrete energy amounts was taking place. As Gearhart has pointed out, Planck repeatedly stressed the need for a physical interpretation of the constants he introduced (2002, 200). In fact, there has been much debate on how Planck actually understood the various derivations he gives of the quantum hypothesis. For instance, Kuhn (1978) among others has argued that Planck was not literally considering quantized energy elements in his 1900; 1901 papers, but was thinking in terms of continuous energy amounts and using the mathematical apparatus of quantized energy as a calculational convenience. This is in opposition to historians such as Klein (1961), who have argued for a more robust understanding of Planck's "energy quanta." Gearhart has provided an overview of the history

and the various interpretive positions, and argues that it is difficult to maintain the view that Planck himself had in mind the quantization of something like phase space as early as 1900 and 1901 (2002). It is worth noting that Planck's own understanding of what exactly is being quantized is not crucial to the point being made here. Regardless of how and why absorption and emission may occur, we can still see how *QH* had unificatory power by examining its application in various phenomena.

3.2 Light Quanta

Although Einstein was aware of Planck's work on blackbody radiation, his own work in radiation theory stemmed from a slightly different motivation and he was in fact reluctant to fully accept Planck's conclusions, as Einstein believed they diverged further from classical theory than Planck himself was aware. His work led him to conclude that "monochromatic radiation of low density ... behaves, in a thermodynamic sense, as it if consisted of mutually independent radiation quanta of magnitude $Rh\nu/kN_0$" ((Einstein, 1905, 143), translation from (ter Haar, 1967, 102)), where I have here replaced Einstein's constant β with the equivalent h/k for ease of reference.

This paper is perhaps best known for Einstein's treatment of the photoelectric effect in producing "cathode rays," or beams of electrons. One instance was the emission of such rays from a metallic surface after the absorption of incident ultraviolet light. This was first observed in 1887 and studied further in subsequent years, particularly by Philipp Lenard. Einstein hypothesized that light quanta penetrating the surface layer of bodies has energy that is transformed into electron kinetic energy within the substance; electrons then escape the surface with a certain kinetic energy having produced some quantity of work. We can consider the equation Einstein describes in terms of our discussion of informational relevance.

The experiments done on the photoelectric effect yield information about the size of h. One can derive a relation between the energy of electrons and the size of h based on the kinetic energy of the electrons being emitted. Einstein reasoned that $\Pi E = Rh\nu/k - P'$, where the body under investigation is charged to positive potential Π, E is the charge of a gram equivalent of an ion, and P' is the potential of negative electricity. Experiments on the photoelectric effect provided observed values for the unknowns in the relation $\Pi E = Rh\nu/k - P'$. Known quantities could then be inserted into this formula: R is a known constant, $E = 9.6 \cdot 10^3, P\prime = 0, \nu = 1.03 \cdot 10^{15}$. ($\nu$ corresponds to frequencies of ultraviolet light, and the other values are given for an experimental setup.) The order of magnitude of Π according to Lenard's results = 10^7. Einstein calculated the theoretical value of ΠE according to his theoretical assumptions, and found that his theoretical value of Π was in good accord with the experimental results of Lenard. This provided a constraint for the value of h even though at the time it could only have been given within an order of magnitude. Because Einstein's β was equivalent to h/k and the order of magnitude of β had 10^{-11}, the measured value of a body's resistance in cases of the photoelectric effect constrained h to be of order of magnitude 10^{-27}.

Let us now explain how this fits into the Bayesian framework by determining how the various experiments provide information about h. First, note that by beginning with ***QH***, one can calculate the average energy of the resonators Planck was considering in order to obtain the radiation formula Equation 4. However, this equation refers only to the form of a family of equations, where the value of h is not yet determined. Thus, let e_1 be the proposition expressing the results of Lummer and Pringsheim's work determining the maximum wavelength of blackbody radiation in air at a given temperature, "$\lambda_m T = 0.294cm \cdot K$." Let e_2 be the proposition that an experiment on the photoelectric effect would yield a result such

that Π is of the order of magnitude 10^7. From e_1, in conjunction with QH as applied in deriving Equation 4, one obtains that the value of $h = 6.55 \cdot 10^{-27} erg \cdot sec$. Similarly, the results of e_2 in conjunction with QH yield the result that h is of the order of magnitude 10^{-27}. Before the suggestion of ***QH***, there was no way to use e_1 to yield information about e_2. Thus, the informational relevance of e_1 to e_2 on background b, given by Equation 1, was very low. After all, there was no way that Lummer & Pringsheim's experiments on blackbody radiation would constrain the behaviour of cathode rays, so $Pr(e_2|e_1\&b)$ should be the same as $Pr(e_2|b)$, thus assigning $I(e_2, e_1|b)$ the value 0. Compare this with the informational relevance value on the assumption of QH along with background b. This is given by the expression $I(e_2, e_1|QH\&b) = Log_2\frac{Pr(e_2|e_1\&QH\&b)}{Pr(p_2|QH\&b)}$. The value of $Pr(p_2|QH\&b)$ is the probability that Lenard's results would obtain, which does not have a particularly high value if considered against a general background. However, once we consider e_1 as well, we can calculate a value for h from the blackbody spectrum, thus constraining the value we would obtain from experiments on the photoelectric effect. This yields a very high value for $Pr(e_2|e_1\&QH\&b)$, arguably a value very close to one, thus making the value of the information relevance of e_1 to e_2 quite high.

Now recall that the unificatory power of QH is given by Equation 2, which measures the difference between the relevance of e_1 to e_2 when including QH in the background knowledge, and excluding it. This nonzero value contributes directly to the degree of confirmation of QH by e_1 and e_2 as measured by Equation 3. Thus, by positing behaviour of radiation in terms of quanta of size $h\nu$, the form of the blackbody radiation spectrum constrained possible values of measurements conducted on the phenomenon of the photoelectric effect by providing information about the size of h.

An interesting point here is that Einstein had different ideas in mind for the understanding of

quantization than Planck; Einstein talked in terms of quantization of light, whereas Planck is somewhat noncommittal. For instance, nine years after his introduction of h, he writes,

> [P]revious electron theories suffer from an essential incompleteness which demands a modification, but how deeply this modification should go into the structure of the theory is a question upon which views are still widely divergent. . . . [Some physicists, including Einstein] even believe that the propogation of electromagnetic waves in a pure vacuum does not occur precisely in accordance with the Maxwellian field equations, but in definite energy quanta $h\nu$. I am of the opinion, on the other hand, that at present it is not necessary to proceed in so revolutionary a manner, and that one may come successfully through by seeking the significance of the energy quantum $h\nu$ solely in the mutual actions with which the resonators influence one another. (1915[1909], 68)

For this reason, a 'common cause' account of the spectrum of blackbody radiation and the various light phenomena mentioned here would be difficult to provide. *QH* itself does not posit any mechanisms that can be understood as causes; quantization might stem from the actions of resonators, or the constitution of light, among other possibilities. Nevertheless, we can see that the phenomena discussed above became relevant to one another on the assumption of even something as general as *QH*, and different interpretations of its 'cause' do not affect its unificatory power.

3.3 Spectral Phenomena

The quantum hypothesis and the quantity h were crucial in early characterizations of the structure of the atom, as well as the behaviour of line spectra, specifically when heated gases

produce lines of different colours. It was observed that the radiation emitted from these heated gases were not of a continuous spectrum as classical mechanics would lead one to expect. Rather, the emitted radiation was of a number of specific frequencies, as manifested in a number of discrete lines on the spectrum. Balmer found a formula describing the emission spectrum of hydrogen gas:

$$\lambda = B\left(\frac{m^2}{m^2 - n^2}\right) \tag{5}$$

where $n = 2$, m is an integer ≥ 2, B is a constant. Written in terms of frequency and explicit values for the constant, and generalized to allow for different integers for n and m, this becomes

$$\nu = \frac{2\pi^2 me^4}{h^3}\left(\frac{1}{\tau_2^2} - \frac{1}{\tau_1^2}\right). \tag{6}$$

However, this formula had no known connection with the other phenomena discussed above. Niels Bohr was able to develop a model of the atom that was able to account for the observed line spectra of different elements, which no other theory had been able to do. On his model, there were set orbits for electrons each associated with set amounts of energy. An electron making the jump from one energy level to another would emit a discrete amount of energy, $h\nu$. Thus, the spectral lines produced by a particular gas when heated corresponded to the differences between discrete energy levels of the electrons moving from one level to another. This explained the observed discrete spectrum. There were problems with this model since it postulated the existence of stationary states, which went against certain laws of classical electrodynamics, but importantly, the preliminary model was able to account for the observed spectrum by incorporating the quantum hypothesis.

Bohr calculated relations between several observable quantities based on Planck's radiation

theory utilizing h; these calculations, with observed quantities, fit with the order of magnitude of h. We can reinterpret this as a way to turn the observed line spectra into information about the size of h: we already knew that Balmer's formula could be used to describe emission spectra. According to Bohr's calculations,

$$\frac{2\pi^2 me^4}{h^3} = 3.1 \cdot 10^{15} \tag{7}$$

The observed value was $3.290 \cdot 10^{15}$.

We can reverse the calculation in order to see how such an experiment would have constrained the value of h. We use the same experimental values that Bohr used for the charge of the electron $e = 4.7 \cdot 10^{-10}$ and the ratio of the charge to mass $e/m = 5.31 \cdot 10^{17}$, as well as the observed value of $3.290 \cdot 10^{15}$ and solve for h in the expression above. The result is $h = 6.38 \cdot 10^{-27}$, which we see is remarkably close to Bohr's previously calculated value. In this way, we see how Balmer's formula carried information about the size of h, which was also given by the blackbody spectrum.

In order to make the informational relevance explicit, let us take e_1, as above, to be the statement of Lummer & Pringsheim's results on the maximum wavelength of blackbody radiation in air, $\lambda_m T = 0.294cm \cdot K$. Let e_2 here be that the constant in Equation 7, in front of the brackets, takes on a value around $3.290 \cdot 10^{15}$. As before, a value of this constant without the assumption of *QH* could a priori have taken on an infinite range of values, and the result of measurements on blackbody radiation would not be expected to be informative about this. Thus, the informational relevance of e_1 to e_2 was low, if not zero. However, by assuming *QH*, the blackbody spectrum provides information about the size of h, thus constraining the possible values that the constant could take. This makes it much more likely that the value of the constant should be the one found (on the reasonable assumption that values close to the

one calculated using Planck's radiation theory would be more likely than those that do not provide numerical agreement). This makes the informational relevance of e_1 to e_2 quite high on the assumption of *QH*, in contrast to its value without the assumption of *QH*. This yields a nonzero value for the unificatory power of *QH* with respect to e_1 and e_2, again contributing directly to the degree of confirmation of *QH* by those phenomena.

After Bohr's success with the hydrogen spectrum, other phenomena related to atomic spectra were used as explicit tests for the value of h. James Franck and Gustav Ludwig Hertz performed experiments on the energy of electrons colliding with molecules of an inert gas or metal vapour (1914[1967]). In particular, their experiments with mercury vapour were able to help determine value of h. Here, electrons of a certain kinetic energy were introduced into mercury vapour. It was known that at relatively high energies, the mercury gas became ionised. However, below this level but at certain energy thresholds, the electrons lost their kinetic energy; this was attributed to inelastic collisions between the free electrons and those bound to mercury atoms. The fact that these only occurred at discrete levels of energy of the introduced electrons was evidence for the idea that the mercury gas atoms could only absorb energy in those discrete quantities. These energy levels corresponded to the observed spectrum lines emitted by mercury gas.

Since the experiment involved only quantities that were pre-determined or measurable such as the energy of the introduced electrons, the voltage drop corresponding to the loss of the electrons' kinetic energy, and the frequency of emitted energy in the spectrum, these results were used to calculate a value for Planck's constant. Franck and Hertz calculated that h had the value $6.59 \cdot 10^{-27}$. An analysis of the informational relevance of this experiment is analogous to the one given above.

3.4 Summary of Informational Relevance

I have presented several phenomena that were unified by the quantum hypothesis, namely, the frequency spectrum of blackbody radiation, light phenomena, atomic spectral phenomena, and the specific heat of diamond. One important feature that I have emphasized is the ability of several of these phenomena to help constrain and measure the numerical value of Planck's constant which was an integral feature of the quantum hypothesis. Below is a table summarizing the values obtained from each of the phenomena discussed above.

Phenomenon	**Value of** h
Blackbody radiation	$6.55 \cdot 10^{-27}$
Light quanta	Order of 10^{-27}
Hydrogen emission spectrum	$6.38 \cdot 10^{-27}$
Mercury gas resonance radiation	$6.59 \cdot 10^{-27}$

These measurements are significant because they demonstrate the idea that various observations, understood in terms of constraining information about a parameter, were able to render previously unrelated phenomena relevant to one another by yielding information implicitly contained in those observations. By increasing the informational relevance of each phenomenon to the other, the unificatory power of *QH* is raised. My previous discussion considered only pairwise informational relevance relations, but the generalization to several phenomena yields the following, taking each of the e's below to represent the results of experiments from the four phenomena listed in the table.

$$U(e_1, e_2, e_3, e_4; QH|b) = I(e_1, e_2, e_3, e_4|QH\&b) - I(e_1, e_2, e_3, e_4|b) \tag{8}$$

Thus, the Bayesian notion of unificatory power of the quantum hypothesis is nonzero, and the degree of confirmation of QH receives support not only from the individual phenomena, but from the fact that QH makes those phenomena relevant to one another.

4 Conclusion

In this paper, I have argued that the type of unification displayed by the old quantum theory can be understood in terms of informational relevance, which yields the result that in a Bayesian confirmational framework, this unificatory power contributed to the confirmation of a quantum hypothesis over and above the evidence taken individually. I have argued that in many of these cases, an account of the mechanisms that would explain the observed behaviour were not available, which makes a causal story for the unification more difficult to provide. While not denying that causal explanations have their place in theoretical justification, I hope to have shown that there is at least one case where even when such unification is not available to us, there is an alternative sense that has epistemic force. Thus, despite the lack of a fully acceptable quantum theory, it was epistemically justified for scientists of the time to pursue the quantum hypothesis.

References

Einstein, Albert. (1905). Ueber einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, *17*, 132–148. Translated by D. ter Haar in *The Old Quantum Theory* 1967, pp. 91–107.

Einstein, Albert. (1907). Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Waerme. *Annalen der Physik*, *22*, 180–190.

Franck, James, and Gustav Hertz, (1914[1967]). On the excitation of the 2536Å mercury resonance line by electron collisions. In *The Old Quantum Theory*, (pp. 160–166). Great Britain: Pergamon Press. Translated by D. ter Haar.

Gearhart, Clayton A. (2002). Planck, the quantum, and the historians. *Physics in Perspective*, *4*(2), 170–215.

Hund, Friedrich. (1974). *The History of Quantum Theory*. Trans. G. Reece. London, Great Britain: George G. Harrap & Co.

Janssen, Michel. (2002). COI stories: Explanation and evidence from Copernicus to Hockney. *Perspectives on Science*, *10*(4), 457–522.

Kitcher, Philip. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher, & W. C. Salmon (Eds.) *Scientific Explanation (Minnesota Studies in the Philosophy of Science, Volume XIII)*, (pp. 410–505). Minneapolis: University of Minnesota Press.

Klein, Martin J. (1961). Max Planck and the beginnings of the quantum theory. *Archive for History of Exact Sciences*, *1*(5), 459–479.

Klein, Martin J. (1965). Einstein, specific heats, and the early quantum theory. *Science*, *148*(3667), 173–180.

Klein, Martin J. (1966). Thermodynamics and quanta in Planck's work. *Physics Today*, *19*(11), 294–302.

Kuhn, Thomas S. (1978). *Black-body Theory and the Quantum Discontinuity: 1894–1912*. New York, United States of America: Oxford University Press, Inc.

Kurlbaum, Ferdinand. (1898). Ueber eine Methode zur Bestimmung der Strahlung in absolutem Maass und die Strahlung des schwarzen Körpers zwischen 0 und 100 Grad. *Annalen der Physik*, *301*, 746–760.

Mehra, Jagdish, and Helmut I. Rechenberg. (1982). *The quantum theory of Planck, Einstein, Bohr and Sommerfeld: Its foundation and the rise of its difficulties, 1900-1925, v. 1, pt. 1 of The historical development of quantum theory*. United States of America: Springer-Verlag New York Inc.

Myrvold, Wayne C. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science*, *70*, 399–423.

Norton, John D. (1987). The logical inconsistency of the old quantum theory of black body radiation. *Philosophy of Science* , *54*, 327–350.

Norton, John D. (1993). The determination of theory by evidence: the case for quantum discontinuity, 1900–1915. *Synthese*, *97*, 1–31.

Norton, John D. (2006). Atoms, entropy, quanta: Einstein's miraculous argument of 1905. *Studies in History and Philosophy of Modern Physics*, *37*, 70–100.

Planck, Max. (1900). Zur theorie des gesetzes der energieverteilung im normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, *2*, 237–245. Translated by D. ter Haar in "The Old Quantum Theory" 1967, pp. 82–90.

Planck, Max. (1901). Ueber das gesetz der energieverteilung im normalspectrum. *Annalen der Physik*, *4*, 553–563. Translated by Koji Ando.

Planck, Max. (1914[1913]). *The Theory of Heat Radiation*. Philadelphia, USA: P. Blakiston's Son & Co., 2 ed. Translated by Morton Masius.

Planck, Max. (1915[1909]). *Eight Lectures on Theoretical Physics: Delivered at Columbia University*. New York: Columbia University Press. Translated by A. P. Wills.

Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, New Jersey: Princeton University Press.

ter Haar, Dirk. (1967). *The Old Quantum Theory: Selected Readings in Physics*. Great Britain: Pergamon Press.

Whewell, William. (1989). *Theory of Scientific Method*. United States of America: Hackett Publishing Company. Ed. Robert E. Butts.

Wimsatt, William C. (1981). Robustness, reliability, and overdetermination. In M. Brewer, & B. Collins (Eds.) *Scientific Inquiry in the Social Sciences*, (pp. 123–162). San Francisco: Jossey-Bass.

Science and Informed, Counterfactual, Democratic Consent

Arnon Keren

Abstract

On many science-related policy questions, the public is unable to make informed decisions, because of its inability to make use of knowledge and information obtained by scientists. Philip Kitcher and James Fishkin have both suggested therefore that on certain science-related issues, public policy should not be decided upon by actual democratic vote, but should instead conform to the public's *Counterfactual Informed Democratic Decision* (CIDD). Indeed, this suggestion underlies Kitcher's specification of an ideal of a well-ordered science. The paper argues that this suggestion misconstrues the normative significance of CIDDs. At most, CIDDs might have epistemic significance, but no authority or legitimizing force.

Word count: 4995 words

1. Introduction

According to a widely held view, democratic deliberations should arrive at results representing collective, informed decisions (Fishkin 2009). According to this view, science must play an important role within democratic decision-making, as the provider of relevant knowledge and information. But the idea that science can play this role faces serious challenges. One type of challenge emerges from the apparent inability of the public to make use of knowledge obtained by scientists. Research has demonstrated the extent of public ignorance on scientific matters, its lack of motivation to engage in serious study and deliberation on such matters, its vulnerability to manipulation, and the extent of systematic attempts to exploit this vulnerability. Thus, what we can reasonably expect to find, is an "underinformed and nondeliberative public" (Fishkin 2009, 7). Accordingly, even if scientists had all the relevant information needed to make a collective informed decision, the public's decision would often not be informed. Call this the problem of *responsiveness to science*.

A different kind of challenge emerges from the fact that scientific activity may not be properly responsive to the values, needs, and interests of different segments of society. The clearest example of this concerns the way in which the scientific agenda is set. Scientists often do not pay enough attention to questions the answer to which would serve the interests of much of the public.[1] Accordingly, even if the public were responsive to knowledge held by scientific experts, available scientific knowledge would often not allow the public to make informed decisions, because the knowledge scientists seek is not that which is relevant to the public and its needs. Call this the problem of *the responsiveness of science*.

While these are distinct problems, a single idea, developed independently by prominent philosophers of science, such as Philip Kitcher (2001) and political theorists, such as James Fishkin

[1] A notable example is the relative lack of attention to diseases which afflict the poor within biomedical research (Flory and Kitcher 2004).

(2009) may seem to suggest a way of addressing them both. The suggestion can be presented via an analogy with the treatment of consent in contemporary medical ethics. In medical contexts, where actual informed consent cannot be obtained because of a patient's lack of decision-making capacity, it is widely held that treatment decisions should be based on what the patient would have decided upon, if he were to make an informed, considered decision. Analogously, when it comes to certain decisions that require scientific input, contemporary democracies may seem to lack the capacity to reach an informed democratic decision. On such questions, it might be suggested, policy should conform not to actual democratic decisions, but to what the public would have decided upon, if it were to reach a decision through an informed, democratic decision-making process. In other words, the problem of responsiveness to science can be addressed, if public policy on science-related issues is not based on an actual democratic decision, but instead conforms to a *Counterfactual Informed Democratic Decision* (henceforth: CIDD).

This suggestion can also be relied upon to address the problem of the responsiveness *of* science. Indeed this is the core idea underlying Kitcher's suggested specification of an ideal of a *well ordered science* (2001; 2011). Thus, Kitcher suggests that in asking what the scientific agenda should be like, or in evaluating contemporary scientific institutions, the standard to which we should appeal is to be specified by ideal, hypothetical democratic deliberations. That is, the actual scientific agenda should conform to the agenda that would be decided upon, if it were decided upon through ideal, counterfactual, informed democratic deliberations.

While I believe that CIDDs can play an important role in our attempt to address the problems of scientific responsiveness, I think that it is a mistake to think that we should make public policy decisions requiring scientific input in accordance with CIDDS, or that CIDDS determine the standards against which actual institutions should be judged. CIDDs, I shall argue, do not have the same normative import, and do not carry the same kind of authority, as either actual democratic decisions or counterfactual decision of incapacitated patients. The latter have legitimizing force,

while the significance of CIDDs is merely epistemic. The idea that public policy should ultimately accord with CIDDS therefore misconstrues their significance.

2. Democratic Decision-Making for Incapacitated Societies

In the medical context, the standard contemporary view is that if a patient lacks decision-making capacity, then, provided certain conditions are met, whoever acts as surrogate decision-maker should "attempt to decide as the patient would have decided in the circumstances that now obtain, if the patient were competent" (Brock 1994, S9). This is often referred to as *The Substituted Judgment Standard* (SJS). Now it might be suggested that when it comes to certain decision, for which scientific input is required, contemporary societies lack decision-making capacity. They are incapacitated, in the sense that they are unable to arrive at collective informed decision, not because relevant knowledge and information is unavailable, but because, like incapacitated patients, they lack the ability to make decisions based on sound deliberation on available information. It is for this reason that the analogy with the treatment of incapacitated patients may suggest itself.

Here is how Fishkin employs the analogy:

> Just as when individuals offer informed consent to a medical…procedure, we think they should know what they are agreeing to…we can apply generally similar considerations to the outlines of an acceptable *collective* process of achieving something analogous—the consent of "we the people." (2009, 34)

Now because our community is in a sense "incapacitated", "[The] choice…is between debilitated but actual opinion, on the one hand, and deliberative and counterfactual opinion, on the other." Fishkin, in advocating the use of deliberative polling, obviously thinks we should go for the latter.

> Deliberative Polling attempts to employ social science to uncover what deliberative public opinion would be on an issue …The resulting deliberative public opinion is both informed and representative. As a result, it is also, almost inevitably, counterfactual. (Fishkin 2009, 26)

It should be noted that Fishkin, while claiming that knowledge of CIDDs of "we the people" has normative import, is not always clear on what its import should be. But the analogy with informed consent in the medical context suggests that under certain conditions, such decision should

be authoritative.[2] And Kitcher, as already noted, defends such a view explicitly with respect to the scientific agenda. Thus, in describing democratic deliberative procedure that should determine the proper scientific agenda, he is explicit that given the ignorance of the public "there's no thought that well-ordered science must *actually institute*" the procedures envisaged. Instead, "[T]he thought is that, however inquiry proceeds, we want it to match the outcomes those complex procedures would achieve…" (Kitcher 2001, 123).

In the medical context, for SJS to be applicable, a number of conditions must hold. First, the decision to be made is one on which the patient's informed decision, if one were actually made, would have been authoritative, but on which the patient has not made a decision; second, the patient must lack decision-making capacity; and third, the surrogate decision-maker must nonetheless be able to know how the patient would have decided, if he were able to make an informed decision. Under these conditions, decision on the patient's behalf should conform to his counterfactual informed decision.

If we take the analogy with the treatment of incapacitated patients seriously, the following might therefore suggest itself. That on questions on which a democratic decision would have been authoritative, but on which no democratic decision has been made; and on which, moreover, our society is unable to actually make an informed democratic decision, we should, as in the medical context, appeal to SJS. On such questions—call them *difficult science-related questions*—we should attempt to decide as the public would have decided, if it were able to make an informed democratic decision.

Of course, one question about the applicability of SJS to democratic decisions is the question whether we can know what decision the public would have made, if it were to make an informed

[2] For a more explicit endorsement of this idea, see, e.g., Fishkin (2002, 234). But often Fishkin does not commit himself to a particular conception of the significance of CIDDs. So my discussion should be understood as adjudicating between different lines of thought found in Fishkin's writings.

democratic decision. However, if we can have counterfactual knowledge about the kind of decisions an individual would have made if he were informed, then there are no principled reasons to think that we cannot have counterfactual knowledge about the democratic decision a community would have made, if it were properly informed. Indeed, social scientists are developing tools whose purpose is precisely to allow us to have such counterfactual knowledge. This is the purpose of *Deliberative Polling*, in which a random, representative sample of the population is invited and incentivized to participate in a deliberative process, culminating in a democratic vote on a policy question that is supposed to represent the decision the entire population would have made, if it could make an informed democratic decision. In recent years such deliberative polls have been put to use to decide on a range of questions—to determine the identity of party candidates in Greece, to guide the energy policy of the state of Texas, to prioritize ways of investment in infrastructure in Wenling city in China, and more (Fishkin 2009).

We can now see how the analogy with informed individual consent might motivate Kitcher's suggestion that the agenda of science should also be set in accordance with a CIDD. Three features of contemporary science seem to support this suggestion. On the one hand, public funding of science and science's significant effect on the public may suggest that an actual democratic decision would have authority in determining the scientific agenda. On the other hand, public ignorance about science may suggest that the scientific agenda should not be determined by actual non-informed democratic decision, but in accordance with a CIDD (Kitcher 2001; 2007).

3. Counterfactual Decisions, Individual and Collective

However, there are reasons for doubt about the suggestion that decisions on the scientific agenda, or on science-related questions more generally, should conform to CIDDs. Indeed, there is a fundamental difference between the normative import of counterfactual decisions of individuals and that of CIDDs that should make us wary of the suggestion.

According to SJS, when we have reliable knowledge about an incapacitated patient's informed counterfactual decision, we should act accordingly, even if we also know that his counterfactual decision would not best serve his interests. We should make decisions based on the best-interest standard only when we have no decisive evidence about how the patient would have decided in this case (Brock 1994). For this reason, if we are certain that a Jehovah Witness, if he were capacitated, would have refused a course of treatment requiring a blood transfusion, we should not provide him with such treatment now that he is incapacitated, even if we are quite certain that such treatment would best serves his interests.

What justifies acting on patients' counterfactual decision even when their decisions are not best? If this is justified, this is because doing so allows us to respect patients' autonomy and right of self-determination by deciding in accordance with their values and conception of the good (Brock 1994). But it is doubtful if a similar justification can be given to the idea that we should make decisions in accordance with CIDDs, even when these decisions do not best serve the community's interest and do not achieve the most just distribution of benefits. There are at least two reasons underlying this doubt.

First, in the case of an incapacitated patient, adhering to SJS when her informed counterfactual decision is not best involves sacrificing the patient's interests in order to respect her own values and conception of the good. In contrast, in the case of a community, acting in accordance with CIDDs when the counterfactual decision is not best—when it does not best serve the interests of community members, or the most just distribution—involves sacrificing the interest of some, in order to respect the values and conception of the good of others. In this sense the analogy studied here, to paraphrase on Rawls, extends to society a principle of choice for one person, thereby failing to take seriously "the distinction between persons" (1971, 27). But this distinction is of prime significance precisely when it comes to the right of self-determination. This right of yours requires

that I respect your values and decisions about how you conduct your own life; but it does not require that I respect your decisions and values when it comes to other persons' lives.

A second objection to the analogy emerges from the observation that unlike the case of an individual, in the case of a community, acting in accordance with the counterfactual informed decision of the majority, and deciding in accordance with the majority's values, principles, and conception of the good may amount to very different things. This is an implication of what has been called the discursive dilemma (Pettit 2001). Consider the following case, of a community composed of three individuals, a, b, and c, each of which has a perfectly consistent set of judgments regarding the truth-value of three propositions (*p*; if *p* then *q*; *q*), as described in table 1:

	p	If *p* then *q*	*q*
a	t	t	t
b	t	f	f
c	f	t	f
majority judgment	t	t	f

Table 1

Here, even though all members of the community have rational sets of judgments, so that all sets include whatever conclusions follow from accepted premises, the same is not true of the resulting collective set of judgments, generated by aggregating individuals' judgments through majority vote. As is well known, this is a general feature of majority vote. Certain principles and facts, from which certain conclusions logically follow, may be accepted by a majority of the population, and yet the conclusions might be rejected by the majority.[3] One implication of this is that while majority vote is a way of making collective decisions that are responsive to individuals' judgments, majority vote cannot be equally responsive to all judgments of all individuals. To be maximally responsive to some judgments it must be unresponsive to others. So the second reasons the analogy with the

[3] This feature of majority voting is not unique to it, and is shared by all aggregation functions that satisfy certain minimal conditions.

incapacitated patient fails is that in the case of a community, deciding on a question in accordance with the counterfactual informed decision of the majority may not amount to deciding on it in accordance with the values and principles of most members of the community.

4. Democratic Decision-Making, Actual and Counterfactual

CIDDs therefore do not have the same normative import as counterfactual informed decisions of incapacitated individuals. The latter are authoritative in the sense that we should abide by them, even if we are quite certain that the decision made is not best. CIDDs do not have this kind of legitimizing force. They can at best serve as an indication of what the best decision is. And they are neither the only relevant kind of indicator, nor the best possible one. Indeed, sometimes the consensus of scientists, or possibly, of philosophers, might provide a much better indication of that.

A supporter of CIDDs-based standards might try to undermine my objections to their authoritativeness by claiming that if my objections were sound, they would not only show that CIDDs should not be authoritative, but that the same is also true of *actual* democratic decisions. And this is surely implausible. However, this counterargument from actual democracy is based on a mistake. The objection to CIDDs-based standards would undermine the authoritativeness of actual democratic decisions only if it equally applied to actual and counterfactual democratic decisions, and only if whatever could be said for the legitimacy of actual democratic decision-making would also apply to CIDDs. But neither of these conditions holds.

Consider the objection from self-determination to the applicability of SJS to the case of the community. One might attempt to undermine the objection by claiming that communities, not only individuals, have a rights to self-government, in virtue of which a community's democratic decision is authoritative, even when the decision is not best, and even if as a result, some individuals' interests are sacrificed. This is something I do not want to contest. However, we need not deny that communities have such rights in order to deny the authoritativeness of CIDDs. For it is a community's actual democratic decisions, not its counterfactual democratic decisions, that are

authoritative. Even if knowledge of hypothetical decisions of agents has normative import, the reason why such knowledge has normative import is not the same as the reason why actual decisions have normative import. As Dworkin noted with respect to Rawls' hypothetical-contract argument for his theory of justice, "A hypothetical contract is not simply a pale form of an actual contract; it is no contract at all" (1975). Actual consent to a contract can generate for me a reason to do what I would otherwise have no reason to do. In contrast, the fact that under certain conditions I would have consented to a contract does not create for me a reason for action. Instead, the force of a hypothetical-contract argument is merely epistemic: it can serve to show what reasons apply to us anyway. Accordingly, if the option that would be accepted by a CIDD is not best, its counterfactual acceptability does not generate for us a reason to accept it.

But then why is that in the case of the individual incapacitated patient, we ought to decide as the patient would have decided, even when we know that his decision is not best? This is presumably so because we have reasons to respect his values, principles and conception of the good, and because his counterfactual decision would follow from these and thus represent them. But it is doubtful whether a community's CIDDs similarly represent its values and principles. In the case of an individual, we normally ascribe counterfactual decision to her on the basis of values, beliefs and principles we know she actually accepts, and we can thus take her counterfactual decision as representing her actual values and principles. But where our grounds for attributing to an individual or collective subject a counterfactual decision is not knowledge of her actual deep commitments, but rather a deliberative poll, or some other such experimental test indicating how she would have decided under counterfactual conditions, it would be a mistake to think that this counterfactual decision represents the subject's actual commitments. Accordingly, even if we can attribute mental

states to collective subjects, these cannot be attributed to them in virtue of merely counterfactual decisions.[4]

The central virtue Fishkin (2009) attributes to deliberative polls is that they arrive at a decision that is both representative and informed. But while deliberative polls may represent the counterfactual informed decision of the public, this counterfactual decision, unlike individual counterfactual decision appealed to in the medical contexts, cannot be assumed to represent the subject's actual values and principles. Therefore, it is not clear why what deliberative polls represent merits special respect. Matters are quite different with actual democratic decisions. Here, by virtue of the actual decision made, we have grounds to attribute commitments to a collective subject. Therefore, my objections to the acceptance of CIDD-based standards do not equally apply to actual and counterfactual democratic decisions.

Consider the second condition that must hold for the counterargument from actual democracy to succeed. Is it the case that whatever can plausibly be said for the legitimacy of actual democratic decision-making similarly supports the adoption of CIDDs-based standards on science-related policy questions? There are reasons to think not. Indeed, there are reasons to think that, quite generally,

[4] If we were to attribute mental states to communities merely on the basis of CIDDS, these would be very bizarre mental states. They would not satisfy minimal conditions of self-knowledge and rationality that any account of mental-state attribution should arguably insist on. Thus, beliefs attributed on the basis of CIDDs would be such that subjects supposedly holding them would normally have no way of knowing that they hold them without performing complicated social-science experiments; nor would such beliefs have any tendency to be expressed in subjects' actual behavior and assertions. And as the discursive dilemma suggests, bodies of beliefs thus attributed to collective subjects would not satisfy condition of rationality, and subjects supposedly holding them would have no way of monitoring their body of belief to ensure that they display properties characteristic rational agency.

plausible arguments for democracy do not support adopting CIDDs-based standards. A detailed argument for this would demand much more space than available here, for arguments for democracy are many and varied.[5] But what has been said above suggests some reasons for thinking so, for it shows that some arguments for democracy, such as arguments from autonomy and self-determination (Christian 2008) do not support adopting CIDDs-based standards. Moreover, instrumental arguments for democracy provide us with further prima-facie reasons for thinking that the objection from actual democracy fails. Such arguments appeal to the good consequences of implementing actual democratic decision-making procedures. For instance, one historically important instrumental argument for democracy is based on claims about the positive effects of participation in democratic decision-making on the character of citizens (Mill 1861/1991). Obviously conforming public policy to CIDDs would not have similar effects on citizens' character. And the same is arguably true of other familiar instrumental arguments for democracy.[6] For according to such arguments, we should insist on the state being governed by actual democratic decision-making procedures, because the good consequences associated with such democratic procedures cannot be obtained otherwise. Hence, if sound, such arguments suggest that these good consequences cannot be obtained by merely conforming public policy to CIDDs. Indeed, if we could obtain the same good consequences in this way, we could get all the good consequences of actual democratic voting and deliberation, while avoiding some of the familiar shortcomings and costs associated with modern democracies. This would constitute a powerful instrumental argument against the claim that we should hold actual, authoritative democratic voting procedures, even when actual democratic decisions can be informed.

[5] Removed for blind review.

[6] Removed for blind review.

5. The Epistemic Role of CIDDs

We thus see that neither the analogy with medical informed consent nor our commitment to democracy should lead us to accept the authoritativeness of CIDDs on science-related questions. Reasons for insisting on the state being governed democratically do not similarly support the claim that public policy should conform to counterfactual democratic decisions. And reasons for giving authority to informed counterfactual decisions of individuals do not support giving authority to informed counterfactual democratic decisions of communities. The analogy with the informed consent of medical patients thus misrepresents the normative significance of CIDDs: The question, which option most people, if informed, would regard as best, is not *the* question we ultimately need to answer when faced with science-related policy issues.

Nonetheless, the answer to this question is significant, and once the misguided idea that public policy should conform to CIDDs is rejected, we can better appreciate its true significance. For the problem with this idea is not only that it attributes to CIDDs a kind of significance that they do not have, but also that it suggests that CIDDs do not have the kind of significance that they should have. If it is the public's *counterfactual* informed democratic decisions with which public policy should conform, then for the well functioning of our democracies, it does not really matter how uninformed the public *actually* is. As long as we are able to know how the public would have decided under ideal counterfactual conditions, it does not matter whether the public learns about science from *Fox News* or from *Scientific American*, or whether our education system provides future citizens with a sound training in science, or in creation science. All that we need to make the right decision on behalf of the public is to know how it would decide, if it were properly informed.

But this is a mistake. Systematic manipulation of evidence made available to the public, and resulting erroneous beliefs held by much of the public, represent a most serious threat to the well-functioning and legitimacy of contemporary democracies. And if consideration of CIDDs does not allow us to avoid this problem, then there is arguably no way of addressing it that does not involve

combating this kind of public ignorance. It is in this context that deliberative polling can have an important *epistemic* role.

Quite generally, knowledge of the opinion of other informed persons is an important epistemic resource. This can be an important epistemic resource even when the informed person is my peer (Christensen 2007). More so, when she is better informed than I am. This is why it is often important for us to know what scientists think on an issue. For this reason, knowing what the public would have judged, if it were informed, offers us with an important epistemic resource. Indeed, as a resource not only for coping with our fallibility and ignorance, but for learning about our fallibility and ignorance, it may have a unique role to play, a role which knowledge of what scientists think cannot play alone.[7]

A developed system of cognitive division of labor would not have created such problems for a democratic decision-making system based on the idea of equal-say to all, if the lay public knew that experts do not differ from the public in their interests, but simply had superior knowledge, and trusted them accordingly. But trust is lacking partially because the public systematically differs from scientists not only in terms of knowledge, but also in terms of interests. So while the encounter with others is an opportunity to learn, both scientists and the public often fail to learn because it is not clear what underlies the difference in opinions between them: different levels of knowledge or different kinds of interests.

It is for this reason that CIDDs may provide both scientists and the public with a helpful mirror. For CIDDs represents the (possible) collective opinion of a group of individuals similar to the public in terms of their interests, but more similar to scientist in terms of their knowledge. A difference between the decision made through a CIDD and actual public opinion may suggest to the public that the difference between actual public opinion and scientific opinion is likely to be the

[7] Moreover, the fact that the collective opinion represented by CIDDs is merely possible, and not actual, makes no difference to its epistemic significance (Kelly 2005).

result of the public's own ignorance. In contrast, if the CIDD is similar to actual public opinion, and different from the opinion of the scientific community, this may suggest to scientists that the differences between their own opinion and that of the public may be explained not by public ignorance, but by something else: perhaps lack of trust, perhaps differences in interests.

Because the division of cognitive labor creates systematic differences between scientists and the lay public both in terms of knowledge and in terms of interests, members of the scientific community can learn from knowledge of the general public opinion, and members of the public can learn from the scientific community's opinion. But members of both communities can learn more from these, if they *also* have knowledge of CIDDs. Devices like deliberative polling might therefore serve an important educational function that neither standard public-opinion polls, nor statements representing the scientific consensus can serve. But to see that this is indeed an important function, we must reject the idea that public policy on science-related questions ought to conform to CIDDs. We must admit that the well functioning of our democracies does not ultimately depend on our knowledge of how the public would decide, if it were properly informed, but rather on how misinformed the public actually is.

References

Brock, Dan W. 1994. "Good Decisionmaking for Incompetent Patients." *Hastings Center Report* 24: S8-S11.

Christensen, David. 2007. "Epistemology of Disagreement: The Good News." *The Philosophical Review* 116: 187-217.

Christiano, Thomas. 2008. "Democracy." In *The Stanford Encyclopedia of Philosophy*, ed. Edward. N. Zalta. Stanford, CA: Stanford University.

Dworkin, Ronald. 1975. "The Original Position." In *Reading Rawls*, ed. Norman Daniels, 16-62. New York: Basic books.

Fishkin, James. 2002. "Deliberative Democracy" In *The Blackwell guide to social and political philosophy*, ed. Robert L. Simon, 221-38. Oxford: Blackwell.

———. 2009. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press.

Flory, James H., and Philip Kitcher. 2004. "Global Health and the Scientific Research Agenda." *Philosophy & Public Affairs* 32 (1): 36-65.

Kelly, Thomas. 2005a. “The Epistemic Significance of Disagreement.” *Oxford Studies in Epistemology* 1: 167-96.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. Oxford: Oxford University Press.

———. 2007. "Scientific Research–Who Should Govern?" *Nanoethics* 1 (3): 177-84.

———. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus Books.

Mill, John Stuart. 1861/1991. *Considerations on Representative Government*. Amherst, NY: Prometheus Books.

Pettit, Philip. 2001. "Deliberative Democracy and the Discursive Dilemma." *Philosophical Issues* 11 (1): 268-99.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Idealization and Structural Explanation in Physics
Draft copy – do not cite

Martin King, 2014

1 Introduction

The focus in the literature on scientific explanation has shifted in recent years towards model-based approaches. The idea that there are simple and true laws of nature has met with objections from philosophers such as Nancy Cartwright (1983) and Paul Teller (2001), and this has made a strictly Hempelian D-N style explanation largely irrelevant to the explanatory practices of science (Hempel & Oppenheim, 1948). Much of science does not involve subsuming particular events under laws of nature. It is increasingly recognized that science across the disciplines is to some degree a patchwork of scientific models, with different methods, strategies, and with varying degrees of successful prediction and explanation. And so accounts of scientific explanation have reflected this change of perspective and model-based approaches have flourished in the explanation literature (Batterman, 2002b; Bokulich, 2008; Craver, 2006; Woodward, 2003).

Of course, not all scientific models are explanatory. Some models are merely calculational tools, whose use in the practice of science is entirely predictive or heuristic, while others are thought to actually explain the phenomena or system they are modelling. The history of scientific explanation in philosophy has focused on articulating independent criteria for what counts as an explanation. In recent work, Alisa Bokulich has argued that idealization has a central role to play in explanation (Bokulich, 2008, 2011, 2012). Bokulich hopes to find a place for certain highly-idealized models to be considered explanatory, even though they are not considered explanatory by causal, mechanistic, or covering law accounts of explanation. She calls these kinds of explanations *structural model explanations* and argues that the structural similarity between the model and the system can debar non-explanatory models (Bokulich, 2008, p. 145). She formulates her account as structural in part to capture models that are not explanatory on Woodward's manipulationist account. She aims to expand the store of explanatory models to include as explanatory those that do not accurately represent, those that model a physical system by means of fictitious entities or processes, what she calls *explanatory fictions*.

The second section of this paper examines Bokulich's account as given in (Bokulich, 2008, 2011, 2012) and articulate her three criteria for explanation. This section will also give an

analysis of Bokulich's argument as it pertains to a case study examined in her book, viz. the phenomenon of quantum wave function scarring. She argues that this very interesting quantum phenomenon is best explained by appropriating concepts and formulae from classical closed orbit theory, rather than by employing quantum mechanical models. This prompts the third section of the paper in which her account is confronted with challenges, in part stemming from a review by Gordon Belot and Lina Jannson, in which they voice concerns over this account's ability to debar non-explanatory models such as Ptolemaic astronomy (Belot & Jannson, 2010). I argue that the structural aspect of her account can in fact debar the Ptolemaic explanation when viewed comparatively, but at the same time it fails to find semiclassical models explanatory. Her own solution to this problem is to use a different aspect of the account to debar Ptolemaic epicycles and allow semiclassical models. However, in section four I argue that this points to a larger worry for structural accounts because the structural criterion is not the deciding factor for which models are explanatory and which are merely phenomenological. Thus on Bokulich's account the measure of structural similarity a model bears to its target system is largely irrelevant to its being explanatory. I conclude with some remarks about what can be learned from Bokulich's work and suggest some ways to move the discussion on explanation forward.

2 Structural Model Explanations

This section examines Bokulich's structural model account of explanation as laid out in (Bokulich, 2008) and incorporates the amendments and clarifications made in (Bokulich, 2011, 2012). Bokulich's account of explanation relies on much of the work done by James Woodward (2003), so the relevant aspects of his account will be briefly recapped first. I then show how this account aims to capture semiclassical models by looking at the phenomenon of quantum wavefunction scarring and demonstrating how it satisfies her account's criteria.

On Woodward's account, causality is framed in terms of counterfactuals rather than in terms of causal mechanisms or physical interaction. Of course not all counterfactuals are going to describe causal relations. He distinguishes between interventionist and non-interventionist counterfactuals. Basically, an explanatory counterfactual tells us what would happen to the systems if certain interventions were to take place. The relations that are invariant under certain changes are doing the work of distinguishing the accidental generalizations from genuine causes. Causal relationships, he claims, are out there in the world, but they are given in the reliable variable dependencies of models. Explanation is the activity of gaining information about these causal relations by discovering through intervention which dependency relations are largely invariant. The counterfactual dependency of these relations gives us important information that provides explanatory depth. This is information that answers what-if-things-had-been-different questions, or *w-questions*. Thus, the range of questions that counterfactual dependence answers is related to the explanatory power of that causal relation (Woodward, 2003; Woodward & Hitchcock, 2003).

Alisa Bokulich adopts aspects of Woodward's account, in particular the idea that giving counterfactual information is central to explanatory power, but she rejects the causal manipulationism. In fact, she aims to give an account that can capture the explanatory power of the structural, non-causal, models that are not captured by Woodward's account. She has in mind the models of semiclassical mechanics. She claims that these models cannot be explanatory on a causal account because the entities involved (electron trajectories) are fictional and have no real

causal power. As it will be shown, the morphologies of the quantum systems of interest depend on the particular periodic orbits of semiclassical mechanics, but the orbits cannot be said to cause the wavefunction distributions, even though there is a reliable dependency relation. Bokulich argues that none of the three main types of accounts of explanation (causal, covering law, and mechanistic) can capture the way semiclassical models explain quantum phenomena, and offers her own *structural model explanation* as a supplement. This account highlights the structural similarities between the real world system and the idealized or fictional model. Bokulich argues that structural model explanations are ones in which there is a pattern of counterfactual dependence among the variables of the model, which can be measured in terms of w-questions, and that this dependence is a consequence of the structural features of the target system (Bokulich, 2008, p. 145).

In developing her account, Bokulich draws on a suggestion made by Margaret Morrison that explanation has to do with structural dependencies (Morrison, 1999). Similar ideas have been developed by John Worrall, James Ladyman, and others (Esfeld & Lam, 2008; French & Ladyman, 2003; Ladyman, 1998; Worrall, 1989). Bokulich offers three general requirements for a structural model explanation, which I have enumerated as follows. The first criterion is E1, which states that the explanation makes reference to a scientific model, *M*. E1 specifies that the explanation is a model explanation and not a covering law or mechanistic explanation. The structural aspect of the structural model explanation comes from the second criterion E2, which says that *M* must be explanatory by showing how there is a pattern of counterfactual dependence of the relevant features of the target system on the structures represented in *M*. E2 is intended to determine which models are genuinely explanatory by ensuring that an explanatory model bears a close structural similarity to the counterfactual structure of the phenomenon. This structural 'isomorphism', as she calls it, is given an objective measure in terms of w-questions (Bokulich, 2008, p. 145). The final criterion is E3, which states that there must be a justification that specifies the domain of the application of *M*. E3 is what she refers to as the *justificatory step*, intended to specify "where and to what extent the model can be treated as an adequate representation of the world" (Bokulich, 2008, p. 146).

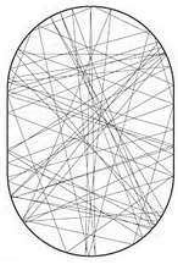
She applies her criteria for explanation to some cases of semiclassical mechanics as part of a larger project of reconceiving the intertheoretic division between the quantum and the classical. She argues that semiclassical mechanics can be genuinely explanatory of certain quantum systems. Semiclassical mechanics is of particular interest to her because they seem to fall outside of the range of other accounts of explanation. The reason seems to be that the models of semiclassical mechanics are non-Galilean, or highly-idealized. The distinction between Galilean and non-Galilean idealizations is one that was made popular by McMullin and it can help clarify the special nature of these models (McMullin, 1985). Very simply, if one can add detail to an idealized model and continually get closer to the real system, then this is what is known as a *Galilean idealization*. For instance, Galileo, in determining the rate of falling bodies, made use of balls rolling down incline planes that were assumed to be frictionless. This kind of idealization is harmless, and the same mathematical relation at which Galileo arrived can be modified to include friction to increase its accuracy. Models featuring these idealizations can be explanatory for McMullin. They represent the target system in a straightforward fashion and their use in explaining the system is justified in part by the fact that they approximately represent. Robert Batterman has described these models as having "controllable" idealizations, in that the idealizations of the system are justified theoretically (Batterman, 2005, p. 235). Idealizations that are non-Galilean on the other hand, have singular limits and cannot be

modified to approach the target system. They are what I refer to as *highly-idealized models*. These models lack the representation that justifies their use in explanation. However, Batterman, Bokulich, and others argue that this does not preclude explanation.

Semiclassical models are prime examples of highly-idealized models because it is not possible to recover the quantum models by adding realistic detail into the semiclassical models. If the semiclassical models have explanatory power, it cannot be due to an underlying causal mechanism of which they are a Galilean idealization. Bokulich thinks that semiclassical models of quantum wavefunction scarring are precisely the kinds of structural explanations that Woodward's 2003 account does not consider explanatory. This is why she allows that the justification of the application of the model to quantum phenomena (E3) can be *top-down* from theory, rather than bottom-up where it would be smoothly recovered in Galilean idealization. For semiclassical mechanics there is no smooth approximation, but there is Gutzwiller's periodic orbit theory. This theory specifies how the classical trajectories can be used to model certain quantum features.

Robert Batterman was the first to argue that semiclassical appeals to classical structures in quantum phenomena at the asymptotic limit between the two is explanatorily important (Batterman, 1992, 2002b). Bokulich claims that structural explanations are actually quite popular in mechanics where appeals to structural restrictions can account for certain aspects of systems. She argues that semiclassical mechanics can be an important interpretive and explanatory tool for certain quantum phenomena, specifically in the subfield of quantum chaos. Classical chaos is found in a great number of systems in which there is an extreme sensitivity to initial conditions, such that immeasurably small differences in initial conditions may result in an exponential divergence. Of course, sensitivity to initial conditions has no part in quantum theory, but quantum models that also describe these systems must exhibit something like chaos themselves. Both Bokulich and Batterman explain that one expects to find a correlate of classical chaos in quantum systems because of the Correspondence Principle, originally formulated by Bohr as the agreement between classical and quantum mechanics as $\hbar \to 0$. The reason is that because classical mechanics is the classical limit of quantum mechanics, there ought to be quantum systems that underlie classically chaotic systems as well (Batterman, 1992, pp. 51-52; Bokulich, 2008). One of Bokulich's strongest examples is that of quantum wavefunction scarring in systems known as *quantum billiard models* (described below). Studies of these quantum billiard systems have revealed that there is an unexpected accumulation of the wave functions along the trajectories that would be periodic orbits in a classical chaotic system.

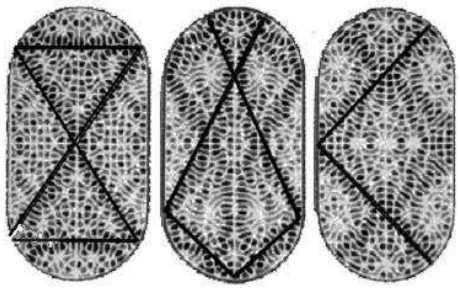
In the classical billiard systems, a stadium shaped enclosed space is inhabited by a free-moving particle whose trajectory is mapped. The shape of the enclosure generally creates a chaotic trajectory that displays an irregular pattern (Fig. 1).This irregular pattern eventually leads to a uniform distribution of trajectories throughout the space. However, there are a few initial conditions that lead to periodic orbits in which the motion of the particle constantly repeats itself. There are certain starting positions and velocities that will not result in a uniformly distributed stadium. This occurs in different shapes including a rectangle, a vee, and a bow tie, among others 0.



(Fig. 1)

Figure 1. A typical example of a classical chaotic trajectory of a particle in a stadium shaped enclosure (Stöckmann, 2010).

In the quantum analog, since no sensitivity to initial conditions plays any part, one would expect to be unable to distinguish the chaotic and periodic orbits. Without orbit theory, there is no reason to expect that anything other than a random superposition of plane waves exhibiting a regular and diffuse pattern. But what one actually finds is that the probability density of the wave functions is significantly higher in certain areas 0. Not all the wavefunctions are evenly distributed. The interesting fact is that they actually converge on the classically stable periodic orbits.



(Fig. 2)

Figure 2. Three eigenstates of quantum billiard stadiums appear to give wavefunction distributions along trajectories predicted by the classical closed orbit theory.

What this suggests is that the shapes of the classically stable orbits in the stadium overlap and make its probability more intense. Gutzwiller's periodic orbit theory is a method of approximating the density of quantum states from classical periodic trajectories by means of the Gutzwiller trace formula (Eq. 1), a semiclassical approximation of Green's function,

(Eq. 1) $$\rho_{osc} \approx \frac{1}{\pi\hbar}\sum_p A_p \cos\left(\frac{S_p}{\hbar} - m_p\frac{\pi}{2}\right)$$

where *S* labels the action of the periodic orbits *p*, *A* is a measure of the orbits' stability and m is the number of times the neighbouring orbits intersect the periodic orbit in one period. Gutzwiller's theory specifies how the behaviour of a Gaussian wavepacket can serve as accurate solutions to the time-dependent Schrödinger equation, and thus how the allowed classical periodic orbits corresponds to the accumulation of wavefunction density observed as the scarring

phenomenon. Bokulich's contention is that the scarring phenomenon is explained better by appealing to the periodic orbit theory than it is by solving the Schrödinger equation.

She argues that classical trajectories, though fictions – false of the quantum system – are explanatorily relevant to the phenomenon of quantum wave function scarring. By falsely assuming that the particle travels along a classical trajectory, one correctly expects to find certain scarring patterns in quantum billiard systems, which one would not expect on a simple quantum picture. She argues that this example is a case of bona fide structural model explanation. This example is not an outlier case, but one of many Bokulich examines, including the conductance peaks of quantum dots, the orbits of Bohr's model of the atom, and the resonance peaks of the Rydberg electrons.

For Bokulich, these examples suggest that there is a "dynamical structural continuity" between the classical and quantum theories, though not as straightforward as relation proposed by Bohr. Because of this she argues that semiclassical fictions, in this case the classical periodic orbit theory applied to a particle, can serve to give counterfactual information about the quantum system. The closed orbits are not real, in the sense that the particle is not actually travelling in a classically defined orbit with position and velocity. Bokulich does not want to argue that the trajectories are real, but rather that they are a special kind of fiction that is also explanatory: "These closed and periodic classical orbits can be said to explain features of the spectral resonances and scarring insofar as they provide a semiclassical model of these phenomena" (Bokulich, 2008, p. 140).

Bokulich admits that it is possible to derive these conductance properties and scarring patterns from a fully quantum picture, by numerically solving the Schrödinger equation, but the dependence of the conductance properties on the classical orbits allows her to say that the semiclassical model is playing an explanatory role here. This dependence conveys physical insight, or structural information, on the quantum dynamics. Of course, in order for Bokulich to claim that there is a genuine explanation here, the semiclassical model must satisfy her three criteria E1-3. And it can be easily shown that they do. The explanation makes reference to a scientific model, viz. Gutzwiller's periodic orbit theory, and so it satisfies E1. The semiclassical mechanics models exhibit the counterfactual dependence of the conductance peaks of the stadiums on the particular classical orbits. It satisfies E2, the structural criterion, because one is able to say how the wavefunction distribution inside the stadium would change if the periodic orbit had been different, or if the shape of the stadium is changed. This putative explanation is also justified in being applied to this domain (E3) because Gutzwiller's periodic orbit theory specifies how to model features of quantum dynamics with classical trajectories.

So for Bokulich, these semiclassical models qualify as explanatory. But Bokulich argues for an even stronger case, viz. that the semiclassical models actually provide *better* explanations than the fully quantum ones. She does not claim that quantum mechanics alone cannot predict these phenomena, but rather that its explanations are deficient because they do not provide as much counterfactual information about the system, which gives us physical insight into the system and grants understanding. In order to get a measure of the information a model gives about the system, she makes use of w-questions and Woodward and Hitchcock's notion of explanatory depth. The more w-questions a model answers, the more information it gives, the deeper the explanation it provides (Bokulich, 2008, p. 152).

For Woodward, the range of w-questions that a model can answer about a phenomenon is directly related to its ability to explain that phenomenon, where models that are more general or more fundamental provide deeper explanations. However, Bokulich claims that the semiclassical

model of wavefunction scarring gives counterfactual information about the quantum system, and further that "there can be situations in which *less* fundamental theories can provide deeper explanations than more fundamental theories" (Bokulich, 2008, p. 153). Given that there are full quantum derivations of these scarring phenomena, if Bokulich wants to argue that the semiclassical model is explanatory and the quantum derivation of the same phenomenon is less so, then she has to show that the semiclassical model can answer a wider range of w-questions.

There would be more room for this argument is it could be argued that the semiclassical models offer answers to a different class of w-question, viz. questions about what the quantum wavefunction scarring would look like if the semiclassical orbit were different, or questions about why these particular morphologies are favoured. Information about why particular scarring patterns, as seen in 0, occur is given by the semiclassical model, so the argument would go, because it is easily capable of accounting for the chaotic and the particular periodic trajectories, and can show how the quantum scarring would change if things (the periodic orbits) had been different.

3 Worries about a Structural Criterion for Explanation

I have shown how Bokulich's account aims to capture the explanatory power of highly-idealized models like those of semiclassical mechanics, and I now turn to examine some worries about these structural model explanations. In the aforementioned review of (Bokulich, 2008), Belot and Jannson are concerned that once the account of structural model explanations allows for such fictions as classical trajectories in quantum systems, it will be unable to reject models that are widely considered non-explanatory, such as those of Ptolemaic astronomy (Belot & Jannson, 2010). The worry is that once she opens the door up to explanatory fictions her criteria are not strong enough to debar non-explanatory fictions, such as planetary epicycles. I shall show that on one reading of Bokulich's account Ptolemaic models can be shown to be non-explanatory. However, the concerns of Belot and Jannson are not misplaced as I shall show that this same reading fails to conclude that semiclassical mechanics are explanatory.

As is well known, the Ptolemaic model of the solar system makes use of epicycles in accounting for the apparent retrograde motion of the planets across the night sky as seen from Earth. Bokulich is explicit in wanting to allow for the idealizations in quantum dots and quantum billiard systems (fictitious electron trajectories), but not those of Ptolemaic astronomy (fictitious epicycles). Belot and Jannson are right to worry that epicycles and electron trajectories are of the same ilk, but Bokulich might have room for admitting one and not the other. At first glance, it seems that the Ptolemaic explanation for planetary motion satisfies her three criteria for a good explanation.

The Ptolemaic explanation satisfies E1, insofar as it references the geocentric model of the solar system. The model is also counterfactually reliable under certain conditions. The Ptolemaic system has trigonometric tables of chords used for calculations, and these give us counterfactual information about the visible solar system. And so it also seems to satisfy E2. It is only on the third criterion E3 that the Ptolemaic models will be debarred according to Bokulich. The geocentric model and its epicycles are not adequate representations of the real structure of solar system, and so not deemed relevant to the explanation of planetary motion by the contemporary state of science (Bokulich, 2012, p. 735). This is indeed true, however, I will return to the third criterion in the following section, but for now I will focus on her assessment of Ptolemaic models and the structural criterion E2.

It is important to remember that E2 is the criterion intended to pick out which models are genuinely explanatory. This is the structural part of structural explanation. The structural isomorphism is meant to replace manipulationist causation in Woodward's picture as the main deciding factor for an explanation. In order to most accurately assess the satisfaction of E2, one needs to actually measure the number of w-answers.

Unfortunately, obtaining a measure of the number w-questions a certain model can answer is not straightforward, and Bokulich gives no real method for obtaining such a measure. The first problem one encounters is in attempting to count individual w-questions. The Ptolemaic system has methods of calculating the positions of the bodies of the solar system for any given day, for any place on Earth, including not just positions in the night sky, but eclipses, solstices, equinoxes, and so on. Importantly, these bodies have cycles and epicycles that are continuous, and so one could get an infinite number of w-question answers, along each of the points on the lines of the spherical trigonometry. And so the number of w-questions the model can answer cannot be meaningfully counted.

Bokulich does not explicitly frame w-questions in a comparative way, but I suggest that Ptolemaic epicycles do answer fewer w-questions (provide fewer w-answers) than the Copernican model, and it can be debarred in that fashion. As we have seen, a quantitative method for counting is not possible, so a simple quantitative comparison cannot be made. However, there is a sense in which an intuitive comparison of the classes of w-questions is possible. Because there is a lot of overlap of the information one gets from the Ptolemaic model and the Copernican model, an argument could be made that the Ptolemaic model has a narrower scope, which is to say that the Copernican model can give all or nearly all the w-answers that the Ptolemaic model provides, but also a lot of additional w-answers as well, such as accounting for the phases of Venus and giving counterfactual information about the positions of the planets when not seen from Earth.

The Copernican model answers more w-questions on this kind of comparison, so perhaps epicycles are not explanatory. With this kind of comparison there seems to be a way after all in which Bokulich can have her structural criterion E2 decide between explanatory and non-explanatory models.

If this comparative framework works for Ptolemaic astronomy, does the same hold in the case of semiclassical mechanics? Well, when one returns to the semiclassical models and attempts to compare the w-information with that provided by quantum mechanics, the comparison does not seem to lead to the conclusion that semiclassical mechanics is explanatory. The semiclassical model can give counterfactual information about the distribution of probability densities in the enclosure in straightforward way. There is a certain range of questions that can be answered about the dependence of the scarring on the classical orbits. It seems that "rather than obscuring the genuine mechanisms at work, this idealization actually brings them into focus" (Batterman, 1992, p. 64). So it seems that there could be a class of w-questions that are better, or more intuitively, answered by the highly idealized model. The highly-idealized model is indifferent to the details and particulars of the dynamics and allows certain features like scarring phenomena to be highlighted.

However, the Schrödinger equation *can* derive the results that are obtained in the semiclassical models, as Bokulich freely admits. But in addition to this, the quantum model can also provide w-answers about many other quantum systems, ones in which the semiclassical model fails to hold. This seems to give the same comparative relation between Ptolemaic and Copernican models of the solar system. The quantum system can be seen to give more w-

answers because it includes the semiclassical and much else. It can be seen that the comparison fails to side in favour of semiclassical mechanics. If there were a domain of phenomena in which the more fundamental theory could not derive the desired results (and I am certain there are many), then the best explanation would be given by the less fundamental theory. In this case however, Bokulich admits that the quantum models *can* account for the scarring phenomena described by the semiclassical models. And so it turns out that even though the classical trajectories can answer interesting w-questions about the particular morphologies of the wavefunction scarring, models from the more fundamental theory will always win out in terms of w-questions when they can account for the same phenomena.

And this is what I believe Woodward and Hitchcock had in mind when they introduced the notion of w-questions. For Woodward and Hitchcock, the models of the more fundamental theory is able to provide more information, to give answers to more w-questions (Woodward & Hitchcock, 2003). For them, this implies that the deeper explanation is given by the more fundamental theory; e.g., General Relativity has more explanatory depth than Newtonian mechanics because it answers a wider range of w-questions. If a theory is more fundamental, then its models can answer a wider range of w-questions. Woodward might accept that these highly-idealized models are explanatory, but less explanatory than fundamental models that offer much deeper explanations, as long as they satisfy his criteria for explanation by exhibiting a degree of invariance under a range of interventions. However, this will not work for Bokulich, because it will not favour the models of semiclassical mechanics over those of quantum mechanics because they have overlapping domains. It is important for Bokulich, not only that semiclassical models be explanatory, but that they actually be better explanations of some quantum phenomena than the fully quantum explanations: "Without knowledge of the classical orbits, our understanding of the quantum spectra and wavefunction morphologies is incomplete" (Bokulich, 2008, p. 154).

A further concern is that this kind of comparison seems only to work in cases where there is overlap in the domain of the phenomena. Where there is no overlap, an intuitive sense of which model answers more w-questions, does not seem to have any bearing on the explanatory power of one theory or the other. If one compares semiclassical mechanics with Ptolemaic astronomy, regardless of how the w-information balance tips, E2 still has no real bearing on whether the models of semiclassical mechanics are themselves explanatory. When there is no overlap in the domain of the models, the comparison is not helpful.

Even if there were a quantitative way to measure the structural similarity using something other than w-questions, this problem persists for Bokulich. Imagine it was possible to give a compressed scalar rating of all the complex representations of structural similarity given by a complicated process of calculations and perhaps insights from measure theory. Now suppose that Ptolemaic epicycles were given a rating of 4, Copernican orbits a 9, and semiclassical mechanical models a generous 12. Even though it received a higher ranking than something like Ptolemaic explanations, it is still reasonable to ask "are the models of semiclassical mechanics explanatory?" And so it does not seem that there can be any way that such a comparative framework can provide a general criterion for explanation. It is only when the domain of the phenomena overlap that this will work. However, because a comparison that ranks semiclassical mechanics as less explanatory than quantum mechanics is inconsistent with her view, this measure will not be helpful for Bokulich.

The main worry for a structural criterion for explanation is that a measure of structural similarity can be given to almost any model, no matter how inaccurately it represents. And so if

one wishes to debar the worst of these then a comparison must be made. However, this comparison, when possible, will always side in the favour of the models of the more fundamental theory and not the highly-idealized model of high-level theory. This does not serve Bokulich's end, but it is not in itself problematic. The general problem is that this comparison is only helpful for models with overlapping domains, and leaves unanswered the question of whether a particular model explains its target phenomenon. Philip Kitcher offers a comparative account of explanation, but it is intended to function irrespective of domains (Kitcher, 1981, 1989). It is also comparative in a winner-take-all fashion, where only the most unifying theory was explanatory. The winner-take-all aspect would be problematic for Bokulich because it would not favour quantum mechanics, but this is not in itself a problem, nor a problem for Kitcher, though his purely syntactic approach to explanation theory choice has its other downsides, which willnot be covered here.

4 Worries about the Justificatory Step

Bokulich's own solution is to debar Ptolemaic epicycles, not with the structural criterion E2, but with the justificatory criterion E3. And so in this section I will analyze this aspect of Bokulich's account and raise some concerns about it playing the major role in distinguishing explanatory from non-explanatory fictions.

This justificatory step has three aspects, which Bokulich has expanded upon in (Bokulich, 2012, p. 736), and which I have labelled here as J1-3. The first, J1, is that an explanatory model involves a contextual *relevance relation* set by the contemporary state of science, which ensures that scenarios like falling barometers causing storms are simply not even candidate explanations. The justification also involves an articulation of the *domain of applicability* J2, wherein it is an adequate representation of the system. To satisfy this there must be either a top-down or bottom-up justification of the model's use, as I described above (Sec. 2). Lastly, and closely related is J3, a *translation key* of sorts that allows information about the model to be translated into conclusions about the real system. There must be some reason why information gained in the model is applicable to conclusions about the world. For example, Gutzwiller's periodic orbit theory specifies how the trace formula (Eq. 1) is related to the actual observed morphologies in the quantum stadium billiard. E3 taken together is something like the explanatory standards of contemporary science. It ensures that the model makes reference to the right kinds of accepted entities, states, and processes, and that the relation between the model and the real system is not merely accidental.

Bokulich does not appeal to E2 in order to debar Ptolemaic explanation, rather she argues that the models fail to be explanatory because they do not qualify as adequate representations of the solar system in contemporary science. Explanatory fictions "represent real entities, processes, or structures in the world, while [non-explanatory ones] represent nothing at all" (Bokulich, 2012, p. 734). She wants to allow for fictions to be explanatory, but only fictions that count as adequate representation – something that can only be decided by the relevant scientific community.

In the context of these two examples, the Ptolemaic model is non-explanatory because the orbits are not adequately representative of the real structure of planetary motion: "given the relevance relation set by the state of contemporary science, epicycles are irrelevant to the explanation of retrograde motion. This is not simply because they are fictional but, rather, because they fail to be an adequate fictional representation of the real structure of our solar

system," whereas "the classical periodic orbits are able to capture, in their fictional representation, real features of the quantum dynamics in the dot" (Bokulich, 2012, p. 735). So the adequacy of the fictional representation as determined by the criteria of E3 can debar Ptolemaic epicycles.

In her response to the worry of Belot and Jannson cited above (Sec. 3), she says: "although the range of w-questions that a phenomenological model can answer will typically be more limited, scope alone cannot distinguish between explanatory and phenomenological models." (Bokulich, 2012, p. 733) She offers instead the idea that the current state of scientific knowledge precludes the possibility of Ptolemaic epicycles being counted as explanatory, in the same way that it ought to preclude falling barometers causing storms – neither satisfy J3. It was shown in the previous section that E2 also debars both Ptolemaic epicycles and semiclassical models. Semiclassical models are not so obviously inadequate as to be excluded from the explanatory store, like shadows explaining flagpole heights, and so they satisfy J1. However, for J2 and J3, the semiclassical models of interest employ Gutzwiller's periodic orbit theory to justify their application and provide a means of getting real-world information from the fictional model. And so the semiclassical models seem to satisfy E3 as a whole and are justified in being used in these systems exhibiting quantum chaos, even though it was shown that they did not satisfy E2. Bokulich has provided a reason for thinking that her account can debar this kind of standard counterexample. The Ptolemaic model is simply not a candidate explanation to begin with, because the fictions it employs are too empirically inadequate for them to be considered representations of the structure of the system.

In the remainder of this section I will raise three worries about E3 and about this kind of criterion as the main deciding factor for explanation. The first worry is that even though she insists that electron trajectories in semiclassical models capture real features of the systems dynamics and Ptolemaic epicycles do not, it is not clear that in distinction from epicycles, classical electron trajectories *are* representative of the true electron dynamics, of the *real* structure of the quantum systems, as she claims (Bokulich, 2012). Part of the requirements of E3 is that entities and processes of the model are considered by scientists to be potentially relevant to the explanation (J1). Earlier, I conceded that the semiclassical models should not be dismissed from potential explanations outright, but this does not imply the positive claim that they do capture real quantum structures. Consider the fact that the predictive success of semiclassical models is rather unexpected. This is so precisely because they are not true descriptions of the systems. It may be that there is a certain range of counterfactual information about the systems' morphologies that can be gathered, but it is not readily understood why it is that the dependency relation holds. Given only the full semiclassical explanation, it is still a bit mysterious why the *quantum* effect would be dependent on the *classical* trajectory. However, if one were able to derive this phenomena and render it expectable on a fully quantum picture, that mystery would disappear. This seems to suggest that the *real* structure of the system is only given in a fully quantum picture, in the same way that the numerical coincidences of Ptolemaic calculations are revealed by more fundamental theories.

The second worry is that because this is supposedly a structural explanation, a lot should depend on the satisfaction or degree of satisfaction of E2, but this does not seem to be the case. E2 is not capable of doing the work of distinguishing explanatory from non-explanatory fictions in the way that Bokulich wants, since it debarred both Ptolemaic models and those semiclassical mechanics in favour of models with broader scopes from more fundamental theories. Due to this, E3 has to do most of the heavy lifting. However, if E3 is largely responsible for maintaining a

threshold for explanation, then there is not much of a sense in which these 3 criteria taken together are independent criteria for explanation. The deciding factor is what satisfies E3, i.e. what is consistent with that currently considered to be explanatory in science, and not with the structurally analogous models. The structural criterion that was intended to pick out which models were genuinely explanatory by showing whether the models exhibited the relevant structural properties of the system failed to do so. In order to make a strong case that semiclassical mechanics can provide *structural* model explanations, the structure that allegedly links the models to the real-world system should determine that.

The last related concern is not only that E2 should distinguish explanatory from non-explanatory in a structural explanation, but that E3 is too context sensitive to do this. It seems as though E3 could be determined, or estimated, with structural information. If one wanted to assess the adequacy of a model's depiction of reality, to determine whether its relation was numerological or correlational and know if the model's information is applicable to the real world system, then its ability to give a wide range of reliable counterfactual information about that system seems a reasonable measure. This information is something that the model can provide on Woodward's account, because it is explicitly manipulationist. But because Bokulich does away with the causal interventions and only imports the notion of explanatory depth, this must be added on as a separate criterion and loses objectivity. On Bokulich's account, there can be no interventions to separate the correlational from the causal, instead it falls on the scientific community to decide if it is adequate. E3 is not meant to employ the measure of w-questions – it is not a measure of structural similarity, but a criterion for ensuring that the model is not known to be phenomenological. The criterion is context sensitive and particular to the details of the model and the current views in science regarding what explains and what accurately represents. What counts as an adequate fictional representation (J1) is a moving target, and may or may not be unanimously agreed on across a discipline.

Even if what represents and what explains were widely agreed upon, there is something missing in this kind of justification – a degree of normativity. When Bokulich argues for the explanatory power of semiclassical mechanics she concludes from the work of Wintgen, Richter, and Tanner (1992), as well as others, that it is more than a tool or a method for generating more simply reliable predictions. Bokulich cites physicists as saying that semiclassical descriptions are desirable because the full quantum-mechanical calculations are cumbersome and elaborate and that the "simple interpretation of classical and semiclassical methods assists in illuminating the structure of solutions" (Wintgen et al., 1992, p. 19). It is in getting the structure of solutions that the semiclassical methods are most useful, i.e. they have much heuristic value. Batterman has argued along similar lines citing the work of W.H. Miller: "Semiclassical theory plays an interpretive role; that is, it provides an understanding of the nature of quantum effects in chemical phenomena, such as interference effects in product state distributions and tunnelling corrections to rate constants for chemical reactions" (Miller, 1986). While these quotations are clearly in favour of the value, and explanatory value, of semiclassical mechanics, it is important to remember that the scientists are unlikely to have in mind a rigorous and philosophically robust notion explanation, complete with independent criteria. And that even if some scientists, or even a majority, do find these models to be explanatory, there is more to a philosophical account of explanation than merely capturing that. An account of explanation should not be merely descriptive, but provide independent criteria capable of assessing putative explanations.

Traditionally accounts of explanation have tried to provide a bar above which certain models are counted as explanatory. (Hempel & Oppenheim, 1948). Woodward also seemed

sensitive to this, particularly when he provided motivation for his manipulability account (Woodward, 2003, p. 93). It is not enough to describe the accepted use of causation (or in this case explanation) without providing sufficient motivation for why that particular conception should adopted.

5 Conclusions and Suggestions for Moving Forward

Semiclassical mechanics is still a very fruitful research avenue, and it is intuitively powerful. It allows us to picture and grasp systems that we should not be able to picture, and frame them in familiar terms. And quite astonishingly, it can give us simple and reliable counterfactual information about certain quantum systems. Semiclassical mechanics is certainly more relevant to the current state of science than Ptolemaic epicycles, because its models are heuristically valuable in providing frameworks for investigating and calculating quantum systems. Bokulich's work on explanation and highly-idealized models is largely connected with her larger project of reconceiving the quantum to classical transition. She has, in explicit detail, gone over cases that cast doubt on a simple reductive picture. The concerns remaining for a structural account should not diminish the contributions she makes to our understanding of Bohr's Correspondence Principle and the intricacies of the quantum to classical transition.

Bokulich has taken bold steps forward in offering an objective measure for determining structural similarity in terms of w-questions. However, this measure proves difficult to determine. I have argued that because an independent, objective measure of structural similarity cannot be made, that an objective comparison can also not be made. I further argued that an intuitive comparison is no help for Bokulich. It can be made, but it always sides in favour of the more fundamental explanation and not the highly-idealized model, thus ruling out semiclassical mechanics, and other minimal models. I was able to show that the epicycles of Ptolemaic astronomy need not be considered explanatory, but the worry then becomes that semiclassical models, because they give less w-information than quantum models, are also not explanatory. Further, this distinction is of no use when the domains of the models are completely distinct. The remaining problem for structural accounts is that even if an objective measure were possible, it would still give no information about whether a model is explanatory across domains or independently. Therefore, I argue that structural similarity cannot distinguish between explanatory and non-explanatory fictions.

Because of this, the other criteria in Bokulich's account had to do the heavy lifting with regards to drawing a line between explanatory and non-explanatory models. But these criteria alone seem only to reflect what is currently thought about whether a model is considered explanatory, and do not give independent reasons to conclude that a model is explanatory. The strong role her third criterion plays is also worrying, not only because the structural aspect of the structural explanation is downplayed, but because it takes away the normative aspect that an account of explanation ought to have, and has traditionally aimed for.

There are many lessons we can take away from the new direction this account has taken and the problems that remain. Highly idealized models are common in science and as other have argued (Batterman, 2002a; Batterman & Rice, 2014; Wayne, 2011), there is reason to consider them explanatory. Stepping out of the shadow of Woodward and expanding the scope of explanation is a next major step in the philosophy of science. Bokulich tries to do so by providing a quantitative measure for structural similarity but it ends up not working in her favour. The purely structural criterion is not helpful in distinguishing explanatory from non-

explanatory idealizations, but more than that, an account of explanation should continue the tradition of offering independent normative criteria for explanation, and be descriptive of, and critical of, the explanatory practices of science.

If Bokulich is correct about the limits of causal accounts and the explanatory virtue of highly-idealized models, and there are good reasons to think that she is, then developing an extended account of explanation and idealization is a worthy aim, and she has contributed a great deal to that end.

References:

Batterman, Robert W. (1992). Quantum Chaos and Semiclassical Mechanics. *Proceedings of the Biennial Meetings of the Philosophy of Science Association, 1992*(2), 15.

Batterman, Robert W. (2002a). Asymptotics and the Role of Minimal Models. *British Journal for the Philosophy of Science, 53*, 17.

Batterman, Robert W. (2002b). *The Devil in the Details*. Oxford: Oxford University Press,.

Batterman, Robert W. (2005). Critical Phenomena and Breaking Drops: Infinite Idealizations in Physics. *Studies in History and Philosophy of Modern Physics, 36*, 225-244.

Batterman, Robert W., & Rice, Collin C. (2014). Minimal Model Explanations. *Philosophy of Science, 81*(3), 349-376. doi: 10.1086/676677

Belot, Gordon, & Jannson, Lina. (2010). Review of Reexamining the Quantum-Clasical Relation: Beyond Reductionism and Pluralism, by A. Bokulich. *Studies in History and Philosophy of Modern Physics, 41*, 3.

Bokulich, Alisa. (2008). *Reexamining the Quantum-Classical Relation: Beyond Reductionsm and Pluralism* New York: Cambridge University Press.

Bokulich, Alisa. (2011). How Scientific Models Can Explain. *Synthese, 180*(1), 13.

Bokulich, Alisa. (2012). Distinguishing Explanatory from Nonexplanatory Fictions. *Philosophy of Science, 79*(5), 725-737.

Cartwright, Nancy. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Craver, Carl. (2006). Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential. *Philosophy of Science, 75*(5), 11.

Esfeld, Michael, & Lam, Vincent. (2008). Moderate structural realism about space-time. *Synthese, 160*, 27-46.

French, S., & Ladyman, James. (2003). Remodelling structural realism: Quantum physics and the metaphysics of structure. *Synthese, 136*, 31-56.

Hempel, Carl G., & Oppenheim, Paul. (1948). Studies in the Logic of Explanation. *Philosophy of Science, 15*(2), 40.

Kitcher, Philip. (1981). Explanatory unification. *Philosophy of Science, 48*.

Kitcher, Philip. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. Salmon (Eds.), *Scientific Explanation*. Minneapolis: University of Minnesota Press.

Ladyman, James. (1998). What is structural realism? *Studies in History and Philosophy of Modern Science, 29*, 409-424.

McMullin, Ernan. (1985). Galilean Idealization. *Studies in History and Philosophy of Science, 16*(3), 26.

Miller, W. H. (1986). Semiclassical Methods in Chhemical Physics. *Science, 23*, 171-177.

Morrison, Margaret. (1999). Models as Autonomous Agents. In M. Morrison & M. Morgan (Eds.), *Models as Mediators: Perspectives on Natural and Social Science* (pp. 38 - 65). Cambridge: Cumbridge University Press.

Stöckmann, Hans-Jürgen (2010). Stoe Billiards. In stoe_billiards.jpeg (Ed.). Sholarpedia.

Wayne, Andrew. (2011). Extending the Scope of Explanatory idealization. *Philosophy of Science, 78*(5), 11.

Wintgen, Dieter, Richter, K., & Tanner, G. (1992). The semiclassical helium atom. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 2*(1), 19-33.

Woodward, James. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, James, & Hitchcock, Christopher. (2003). Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Nous, 37*(2), 181-199.

Worrall, John. (1989). Structural Realism: The Best of Both Worlds? *Dialectica, 43*(1-2).

0

Against Lawton's contingency thesis, or, why the reported demise of community ecology is greatly exaggerated.

Stefan Linquist[1]

Abstract

Lawton's contingency thesis (CT) states that there are no useful generalizations ("laws") at the level of ecological communities because these systems are especially prone to contingent historical events. I argue that this influential thesis has been grounded on the wrong kind of evidence. CT is best understood in Woodward's (2010) terms as a claim about the instability of certain causal dependencies across different background conditions. A recent distinction between evolution and ecology reveals what an adequate test of Lawton's thesis would look like. To date, CT remains untested. But developments in genome and molecular ecology point in a promising direction.

Word count= 4,899

[1] Department of Philosophy, University of Guelph.

1. Introduction

Ecologist J.H. Lawton has developed one of the most influential recent critiques of community ecology (Lawton 1999). His discussion is framed around the question of whether community ecology admits of "general laws." This branch of ecology studies multi-species assemblages. It thus focuses on a level of biological organization above (single species) populations but below entire ecosystems. Lawton argued that there are no "useful generalizations," or laws, at the community level as such. His reason is that communities are subject to a wide range of *contingencies* that make it impossible to generalize from one instance to the next. For example, particular communities are shaped by different geological events. They each receive a different pool of migrants in a particular order. They experience different patterns of fire, flood, storm, and so on. These one-off events can dramatically impact the composition of a community. Hence, Lawton proposes that the rules governing community composition are transitory and idiosyncratic. However, he thinks that there is hope for generality at other levels of ecological investigation. Law-like regularities obtain at the (lower) population level and at the (more inclusive) macroecological level. They are found at the population level, according to Lawton, because these systems are simpler and behave in a more uniform fashion. By contrast, at the macroecological level regional contingences become less influential. At this level, one looks at ecosystems on a broad geographic and temporal scale, "whereby a kind of statistical order emerges from the scrum" (1999, 183). These considerations inspired Lawton to pronounce the end of community ecology as a viable discipline:

> In sum, community ecology may have the worst of all worlds. It is more complicated than population dynamics, so contingent theory does not work, or rather, the contingency is itself too complicated to be useful. But paradoxically, community ecology is not big and bold enough to break out of the overwhelming complexity within which it appears to

> be enmeshed. All this begs the question of why ecologists continue to devote so much time and effort to traditional studies in community ecology. In my view, the time has come to move on. (ibid)

Many ecologists have heeded this suggestion. Lawton's paper has received an average of 37 citations per year since its publication, mostly endorsing his contingency thesis. Others who view their research as significant beyond the local field or stream find Lawton's conclusion unbearably pessimistic (Chave 2013). These community ecologists soldier on in the search for generality despite Lawton's warnings. Here I argue that they are correct to do so.

Lawton's argument assumes that if a community has been influenced by unique historical events, then it cannot be explained in terms of law-like processes. I argue that this assumption is confused about the explanatory roles of ecology and history. Specifically, it views these two types of explanation as mutually exclusive. An alternative picture has recently been developed in the field of genome ecology (Linquist et al. 2013), which can be applied to the community level. On this picture, ecological and evolutionary (or historical) explanations make different idealizations about the same underlying process. In its pure form, ecological explanation treats focal entities (genes, populations, communities, etc.) as static types while focusing on how their intrinsic properties interact with features of the environment. Evolutionary explanation, in its pure form, accounts for changes in focal entities over time while ignoring relations to the environment. More will be said, momentarily, about these two modes of explanation and how they are sometimes used conjointly –such as in explanations of evolution by natural selection. The important thing to note is how this picture refutes Lawton's argument. A useful analogy can be drawn to the field of developmental biology. In this field, a purely genetic explanation attempts to idealize over environmental differences, while a purely environmental explanation ignores genetic differences. The field of genetics has moved beyond the simple-minded idea that

evidence of an environmental influence negates the possibility of a genetic explanation. Rather, geneticists have developed statistical techniques for determining, given a certain pattern of variation in some trait, how much of it is explained by genetic and environmental factors, respectively. The same approach applies to historical and ecological factors and their influence on community composition. When we adopt this approach, it becomes apparent that Lawton's contingency thesis is based on the wrong kind of data. He argues from evidence of contingency in particular communities to the conclusion that patterns of variation among communities cannot be explained by ecological laws. This would be like inferring from evidence of a genetic influence on some trait, in a particular individual, that variation among individuals cannot be explained by environmental factors. In both cases the reasoning is fallacious. Thus, it remains an open question whether ecological communities can be explained in terms of law-like relations to the environment.

This essay will proceed as follows. Section 2 offers a more precise statement of Lawton's contingency thesis by drawing on Woodward's (2010) concepts of stability and contingency. Section 3 reviews Lawton's evidence for the contingency thesis. Section 4 introduces the operative distinction between ecology and evolution (or history). Section 5 applies this distinction to the community level and explains why Lawton's evidence falls short of supporting his conclusion.

2. Interpreting Lawton.

Philosophers and ecologists disagree about the conditions for natural laws (Colyvan 2003; Lange 2005). My current aim is not to wade into these disputes. Instead, I offer an interpretation of what Lawton means by "contingent" and how to best define the field of community ecology.

Clarifying these terms is a necessary first step in understanding his argument that there are no stable (or law-like) generalizations in this field.

Lawton distinguishes laws from patterns on the grounds that, "Patterns are regularities in what we observe in nature; that is, they are 'widely observable tendencies;'" whereas laws are the "general principles that underpin and create the patterns" (1999, 178). This statement suggests that Lawton views laws as causal generalizations, while patterns are mere correlations. Lawton notes that patterns can vary in their generality: "Indeed they raise the vexing problem of how many exceptions to general patterns might exist before we would no longer regard them as patterns" (ibid). A similar problem arises for laws regarding their generality. Although some interpretations of Lawton take him to view laws as universal or exceptionless (Roughgarden 2009), this would render Lawton's position rather uninteresting. Exceptions are found even in the laws of chemistry and physics (Cartwright 1983). Hence it would be no surprise to find exceptions in ecological laws also.

Lawton's position is better stated using philosopher James Woodward's concepts of causal stability and contingency (Woodward, 2010). For Woodward, causal relations are represented as counterfactual dependencies among variables. Thus, some variable Y is counterfactually dependent on another variable X just in case, for some set of background conditions B, an intervention that changes only the value of X will result in a corresponding change to Y. The stability (or contingency) of a dependency is defined by the range of background conditions (B) across which it obtains. Thus, a highly stable (for current purposes, law-like) relationship between X and Y is one that holds across a wide range of background conditions. Contingency is the opposite of stability, where a dependency is restricted to a limited range of background conditions.

These ideas are easily transferred to community ecology. Typical dependent variables (Y) in this field include species richness, average abundance, or trophic structure of a community. These are ensemble properties of multi-species assemblages. Typical independent variables (X) include the abundance of a general predator, degree of niche overlap, or other factors thought to impact a community. Background conditions (B) come in at least two dimensions: taxonomic distance (e.g. different phyla or families) and habitat type (e.g. aquatic, marine, and terrestrial habitats). Thus, in some communities it has been observed that increasing the abundance of the top predator increases the diversity of shared prey. This causal relation is stable, in Woodward's sense, to the extent that it holds true for different taxa or across different habitats. A contingent ecological dependency is one that holds for few taxa or habitat types.

Lawton defines community ecology as the study of sets of coexisting species interacting at local scales. This discipline is distinct from population ecology, he claims, insofar as community ecologists study assemblages greater than just two or three species. Although some community ecologists might object to this restriction on their discipline, it is not an issue that I consider here. However, I do take exception to Lawton's requirement that community ecology studies only *local* interactions. The question of how to circumscribe communities as objects of study remains a challenging issue (Sterelny 2006). Lawton suggests that community ecology restricts its focus only to local interactions, so that processes like immigration, emigration, or other meta-community dynamics fall under the purview of macroecology. This will strike many as an artificial way to distinguish these disciplines. Community ecologists should be allowed to circumscribe the boundaries of their subject matter as they see fit and as nature dictates.

Instead of drawing the community/macroecology distinction in terms of local/non-local interactions, a more useful distinction is drawn between the kinds of processes that these

disciplines investigate. Community ecologists have traditionally set aside questions about long term evolutionary processes, focusing instead on the relatively short term processes governing the abundance and distribution of species. Strategically, this simplification makes sense if it indeed turns out that evolutionary processes have only a marginal influence on community composition and abundance. Community ecologists also tend to ignore changes in community composition considered over geological time scales. Over such extended periods, community composition and dynamics are expected to vary considerably (Kricher 1998). By contrast, the macroecological perspective, which Lawton favours, takes both evolutionary and historical processes into account. As Lawton explains, "macroecology is a blend of ecology, biogeography, and evolution and seeks to get above the mind-boggling details of local community assembly to find a bigger picture" (1999, 183). My suggestion is simply that the distinction between community ecology and macroecology is best drawn by focussing on the kinds of process that these disciplines investigate. Community ecology ignores, as a simplifying assumption, evolutionary and historical changes in the focal entities that it investigates; while macroecology attempts to incorporate those changes as well as the events and processes that generate them. This way of drawing the distinction avoids thorny issues about how to draw the boundaries around a community or what constitutes a "local" scale.

To summarize my interpretation of Lawton's position: the counterfactual dependency relations identified for species assemblages greater than 2-3 members are unstable (contingent) across different background conditions such as distinct taxa and habitats. But contingency is reduced either by dropping down to the population level, or, by taking into account broad evolutionary or geological times scales. I refer to this as the *contingency thesis*.

3. Evidence for the Contingency Thesis.

Lawton's central piece of evidence in support of the contingency thesis is based on his 20 years researching a particular bracken fern community located in Skipwith, England. He explains that the relative abundances of these 17 insect species were highly predictable over short (multi-year) time periods – rare species stay rare and more common ones stay common. He adds that the composition of the community is strongly constrained by a species of predatory ant. From Woodward's perspective we can think of this as an invariance relation in which abundance of the predator (X) influences composition and relative abundances of the other members of the insect community (Y). However, Lawton suggests that this relationship is not stable across different background conditions (B).

> I observed an average of about 17 herbivorous insects feeding on bracken at Skipwith each year. Why 17? In crude order of magnitude terms, why not 2? Or 170? This most basic aspect of community structure may have surprisingly little to do with the local processes that dominate so much of traditional thinking in community ecology. (1999, 184)

Lawton goes on to identify two different types of "filter" that, he thinks, determine community composition to a greater degree than those considered by community ecologists. The first is a historical or evolutionary filter: "understanding the origins of the pool requires a knowledge of the evolutionary history of the biota, of geology, of plate tectonics, and so on" (ibid). He suggests, for example, that if members of this community had arrived in a different order it would have altered the relationship between predator and prey abundances. Lawton's suggestion is that any number of one-off events could have significantly impacted community dynamics. Since historical events presumably differ from one community to the next, he reasons, different communities will not obey the same causal dependencies.

The second sort of filter that Lawton identifies is spatial. He proposes that local community dynamics are often influenced by such factors as their distance from a source of migration or overall meta-community structure. Lawton seems to be relying here on the aforementioned stipulation that communities are essentially local. In the previous section I argued that community ecologists are not required to restrict their focus to local species assemblages (whatever that might turn out to mean). Rather, they are free to expand or contract their field of investigation as the situation demands. Thus, if Lawton thought that the composition of his bracken fern community was largely influenced by immigration from another community down the road, he might just have considered them together as a single unit. Lawton distinguishes community from macroecology in such a way that the former is limited both temporally and spatially in its purview. I argue that the field does in fact take on a different character when historical and evolutionary considerations are taken into account. But it is less committed to a particular spatial scale. Hence, we can restrict our focus to the first of Lawton's two filters and ask whether a science of ecological communities can find generality while ignoring historical and evolutionary considerations.

4. Distinguishing Evolution from Ecology.

What then is the relationship between ecology and history? For that matter, what makes a generalization *ecological* in the first place? A candidate solution to these questions has recently emerged within the field of genome ecology (Linquist et al. 2013). This burgeoning sub-discipline applies ecological thinking at the level of the genome, viewing families of mobile genetic elements as akin to species and stable features of the genome as the environment (Brookfield, 2005). As is often the case, applying a familiar theory to a novel domain requires

close attention to its core commitments. This has led to the following operational definitions of "evolution" and "ecology."

1) A strictly evolutionary approach investigates change (or the lack thereof) in some focal entity over successive generations without taking into account its relationships to particular features of the environment.
2) A strictly ecological approach assumes (for simplicity) no change in the focal entities themselves, but focuses instead on the relationships between those entities and features of their environment.

In the following section I apply these definitions to the community level and explain how a strictly evolutionary approach is equivalent to what Lawton would classify as an historical approach. The remainder of this section explicates this distinction and shows how it can be used to determine the extent to which some patterns calls for an ecological or evolutionary explanation.

It is important to note that each mode of investigation is being defined here in its "strict" or pure form. This is just to say that, considered on its own, each approach makes different sorts of idealizing assumptions. For example, the work of Michael Lynch (2007) exemplifies of a purely evolutionary approach. His "mutation hazard" model proposes that large amounts of genome evolution can be explained just in terms of mutation rate (M) and effective population size (Ne). The focal entity in this case is a population or gene pool. M and Ne are variables that apply to intrinsic features of a gene pool, they ignore its relations to features of the environment. In particular, natural selection is not taken into account by this model. It is assumed that when Ne is low the influence of selection on genome evolution is negligible. This is just to say that the environment is ignored by this model under certain conditions. Suppose, then, that the dependent variable of interest is the degree of genetic divergence among a range of related species. Lynch

might explain this pattern of variation by appealing to mutation rate, the length of time over which the populations have been isolated, and the respective population sizes. This would qualify as a strictly evolutionary explanation according to definition 1, since the pattern is being explained in terms of changes in the focal entity while idealizing away relations to particular features of the environment.

Strictly ecological explanations are perhaps even more familiar. Ecologists routinely conduct studies of populations that focus exclusively on their relation to the environment while ignoring changes in the focal entities themselves. For example, the introduction of the Canadian beaver to Argentina in the 1940s led to a population explosion. Here the focal entity is a particular population and the relevant dependent variable is its growth rate. Ecologists attempt to determine which of several possible ecological variables (e.g. lack or predators, suitability of habitat) best explain the much higher rate of population growth in Argentina compared to North America. These studies attempt to account for differences in this dependent variable in terms of various relations to the environment (Anderson et al. 2006). However, they do so without considering whether northern and southern populations differ genetically. That is, they tend not to consider whether there has been change in the focal entities that might accounts for their differential growth rates. Presumably there are good reasons for thinking, in this case, that genetic differences are negligible. The relevant point is that this mode of explanation is purely ecological in that it assumes of focal entities that they are a static type (beavers are beavers, regardless of the population) while focusing on their relation to the environment.

Of course, many patterns in nature cannot be explained either in strictly evolutionary or strictly ecological terms. Often the two types of factor interact. In these cases, it is often necessary to consider how relations between the focal entity and its environment influence

subsequent changes in the entities. This would qualify as a combined or "hybrid" explanation – one that incorporates both evolutionary and ecological factors. Explanations of evolution by natural section are a familiar example (Endler 1986).

Hybrid explanations are undeniably more epistemically demanding than either form of strict explanation. For this reason it is often preferable to establish whether a purely evolutionary or purely ecological model will account for most of the variation in some variable of interest. It is prudent to address this question before attempting to consider both evolutionary and ecological factors in conjunction. There is no need to adopt a more complicated hybrid model if a simpler model will do. Within genome ecology a straightforward strategy has been developed to determine the extent to which a given pattern can be explained by ecological or evolutionary factors (Linquist et al. 2013). One begins with a dependent variable of interest. A population of entities is then selected in which there is variation in the dependent variable. Variation in the dependent variable is required in order to determine the relative contributions of ecological and evolutionary factors. The next step is to identify independent ecological and evolutionary factors that are likely to influence the dependent variable. It is here that definitions 1 & 2 come into play. Evolutionary variables are ones that identify changes in the focal entities over time. For example, in the case of genome ecology, phylogenetic distance is used as a proxy for their evolutionary or historical divergence (ibid). Ecological variables are features of the environment thought to stand in a casual relation to the dependent variable. Admittedly, it is conceptually and empirically challenging to identify independent (ecological and evolutionary) variables that are suitable for this kind of an analysis. Those variables must themselves vary among entities in the sample. Only then can one determine how much of the variation in the dependent variable

correlates with ecological and evolutionary factors, respectively. But once the relevant variables are identified, conducting this type of analysis is a fairly simple matter of statistical regression.

5. Identifying Generality at the Community Level.

Recall that Lawton was worried about the disproportionate influence of historical "filters" on communities. He proposed that various one-off events would dramatically alter their composition and dynamics. We can think of these events as equivalent to the evolutionary factors identified in definition 1. Imagine a community that experiences some unpredictable disruption such as a fire or flood. On the one hand, this might seem to be an "ecological" influence since it is externally imposed on the community. However, by hypothesis these are one-off events. Hence they cannot be treated as *variables* that take on various values across a range of communities. To treat these events in such a fashion would just be to regard them as ordinary ecological factors. To be sure, in some instances fire or flood might be viewed as a quantitative ecological variable. But we are interested here in what it means for these rare events to serve as a historical filter that potentially mitigates an ecological explanation. To view these events as historical contingencies, I suggest, involves viewing them just in terms of their effects on community structure and not, as it were, as types of causes. In other words, when considering the impact of one–off events the relevant question concerns their impact on a community, and not whether the event was a fire, flood, or some other factor per se. Insofar as these events have the same type of effect there is no point in distinguishing them. By analogy, Lynch's model is interested in how changes in Ne impact the fixation of alleles. It doesn't matter about which particular events lead up to a change in Ne. For explanatory purposes these "environmental" factors are treated as a generic kind of cause. Hence the explanation abstracts away from particular relationships to the environment. Much the same applies to the one-off events that Lawton was concerned about.

Let us then consider how definitions 1 and 2 are applied to an ecological community. Suppose that the focal entities are insect communities such as the one Lawton observed. In order to conduct a regression analysis we require a population of these communities that vary in some (quantitative) dependent variable. Following Lawton, let's choose *rank abundance* as the relevant dependent variable. This standard measure in community ecology plots the relative abundance of community members against their rank in abundance, thus generating a curve with a particular shape and slope for each community. The advantage of this as a dependent variable is that it provides a common measure for comparing taxonomically distinct communities.

Lawton's example of predator density is a suitable independent ecological variable, provided that it also varies across the set of communities in the sample. Of course, numerous other ecological variables might be selected. It bears mentioning that there is a considerable danger of false negatives when applying this framework to test for ecological influences on some dependent variable. Unless one selects the correct independent variable, an ecological influence could easily be overlooked.

A greater challenge concerns the selection of historical variables. In the case of genome ecology, phylogenetic relatedness served as a proxy for historical or evolutionary distance. Thus, it was possible to determine how much of the variation among genomes in a sample correlates with phylogenetic distance. The problem is that prototypical communities are less cohesive than genomes. Their members move independently from one community to another. Hence one cannot easily reconstruct a phylogenetic tree for a sample of communities. How then might one identify a quantitative variable to stand in for historical distance?

These limitations are indeed challenging when it comes to most *prototypical* communities. It might simply turn out that assemblages of macro flora and fauna are poor

choices for testing the contingency thesis. However, recent years have seen increased interest in molecular and genome ecology. Diverse communities containing thousands of microorgansims can be contained in a single test tube (Swenson et al. 200), or, in the case of gene families, uploaded to a database. These communities are easily isolated as cohesive units with divergent histories. Thus the molecular and genetic levels offer ample opportunity to test for the influence of chance historical events on community level variables. With this qualification in mind we can imagine how one might test for the stability of an ecological relationship. This would involve comparing the influence of ecological and evolutionary variables across a range of different taxa and habitat types. There are a wide range of molecular and genetic systems in which these experiments could be conducted. Similarly, the dependency between predator abundance and rank abundance could be tested across a range different habitat types. Lawton's contingency thesis would predict little stability in ecological relationships among these different types of community and distinct habitats. To date, no adequate test of this hypothesis has been conducted.

Thinking back to Lawton's argument it becomes clear that he was in no position to pronounce the demise of community ecology. It is a straightforward fallacy to assume that the presence of a historical explanation for some particular community undermines the explanatory power of ecological laws. Nor would it make quantitative sense to ask, "How much of the Skipwith bracken fern community was determined by its historical and ecological factors, respectively?" Any given community will be influenced by both. To partition the relative contributions of ecology and history one must compare a population of communities in which there is variation in the dependent variable of interest. One also requires a way to quantify ecological and historical influences on that dependent variable. Only then, by looking for

ecological correlations that obtain across a range of background conditions, can one determine the stability or contingency of an ecological dependency.

6. Conclusion

Perhaps the take-home message from this discussion is that demonstrating contingency in community ecology is no simple affair. Only certain communities will lend themselves to the kind of quantitative analysis that I have outlined. There are significant challenges associated with identifying and measuring the relevant variables. Even if one finds an apparent influence of history on the dependent variable, there will be looming questions about whether some unidentified ecological variable is perhaps being overlooked. To make matters more complex, an assessment of stability or contingency must proceed across a diverse range of taxa and habitats. In fairness to Lawton, neither the conceptual framework nor the requisite data were available at the time he was writing. However, I have suggested that recent advances in molecular and genome ecology make it easier to test the contingency thesis. As it stands, Lawton's thesis has been supported by the wrong kind of data. It therefore remains an open question whether there are stable ecological generalizations at the community level.

References

Anderson, C. B., et al. (2006), "The effects of invasive North American beavers on riparian plant communities in Cape Horn, Chile. Do exotic beavers engineer differently in subantarctic ecosystems?" *Biological Conservation* 128, 467–474.

Brookfield, John F. (2005), "The ecology of the genome - mobile DNA elements and their hosts", *Nature Reviews Genetics* 6: 128–136.

Cartwright, Nancy (1983), *How the Laws of Physics Lie*. New York: Oxford Press.

Chave, Jerome (2013), "The problem of pattern and scale in ecology: what have we learned in 20 years?" *Ecology Letters* 16(S1): 4-16.

Colyvan, Mark, and Ginzburgh, Lev (2003,) "Laws of nature and laws of ecology", *Oikos* 101: 649-653.

Endler, John (1986), *Natural Selection in the Wild*. New York: Princeton University Press.

Kricher, John (1998), "Nothing endures except change: Ecology's newly emerging paradigm", *Northeastern Naturalist* 5: 165-174.

Lange, Mark (2005), "Ecological laws: What would they be and why would they matter?" *Oikos* 110: 394-403.

Lawton, John H. (1999), "Are there general laws in ecology?" *Oikos* 84: 177-192.

Linquist, S. et al. (2013), "Distinguishing ecological from evolutionary approaches to transposable elements," *Biological Reviews* 88: 573-584.

Lynch, Michael (2011), "Statistical inference on the mechanisms of genome evolution", *PLOS Genetics* 7: 1-4.

Roughgarden, Joan (2009), "Is there a general theory of community ecology?" *Biology & Philosophy* 24: 521-529.

Sober, Elliott (2000), "Appendix one: the meaning of genetic causation." In *From Chance to Choice – Genetics and Justice* (eds. A. Buchanan, D. Brock, N. Daniels and D. Wikler). New York: Cambridge Press, 349–373.

Sterelny, Kim (2006), "Local ecological communities", *Philosophy of Science* 73: 215-231.

Swenson, W. et al. (2000), "Artificial ecosystem selection", *Proceedings of the National Academy of Sciences* 97: 9110-9114.

Woodward, James (2010), "Causation in biology: stability, specificity, and the choice of levels of explanation", *Biology & Philosophy* 25: 287-318.

What the 19th century knew about taxonomy and the 20th forgot

P.D. Magnus

July 6, 2014

For presentation at the *Philosophy of Science Association* biennial meeting in Chicago, Illinois (November 2014).
This is a draft. Comments are welcome.
e-mail: pmagnus(at)fecundity.com
web: http://www.fecundity.com/job

Abstract

The accepted narrative treats John Stuart Mill's Kinds as the historical prototype for our natural kinds, but Mill actually employs two separate notions: Kinds and natural groups. Considering these, along with the accounts of Mill's 19th-century interlocutors, forces us to recognize two distinct questions. First, what marks a natural kind as worthy of inclusion in taxonomy? Second, what exists in the world that makes a category meet that criterion? Mill's two notions offer separate answers to the two questions: natural groups for taxonomy, and Kinds for ontology. This distinction is ignored in many contemporary debates about natural kinds and is obscured by the standard narrative which treats our natural kinds just as a development of Mill's Kinds.

This paper concerns debates about classification in the 19th century between Willaim Whewell (§2), John Stuart Mill (§3), and some lesser known critics (§4). I aim to show that Mill provides separate answers to two important questions in the neighborhood of what we would now call natural kinds: The *taxonomy* question, about what distinguishes categories which are natural kinds from categories which are not; and the *ontology* question, about what there is in the world which sustains that difference. Mill distinguishes *natural groups* as an answer to the taxonomy question and *Kinds* as an answer to the ontology question for some — but importantly not all — natural groups. This overturns the usual story, according to which Mill's Kinds map neatly on to our natural kinds, and it also reveals a distinction we would do well to remember.

1 The standard story

The standard narrative, promulgated by Ian Hacking [4], is that the philosophical conception of *natural kinds* descends from John Stuart Mill's notion of *Kinds*

(with a capital 'K'). According to Hacking, this was a central piece of a promising research program in the mid-19th century which has since degenerated. He argues that the concept is no longer fruitful and so natural kinds should be abandoned. In Hacking's metaphor, Mill's contribution was the "rosy dawn" for natural kinds, present debates are a "scholastic twilight", and the day for thinking in terms of natural kinds has come to an end [5].

Hacking's narrative is widely accepted. For example, John Dupré gestures to the history of natural kinds by writing, "Ian Hacking reminded us that the contemporary tradition of natural kinds arose... in the nineteenth century..." [2, p. vii]. The story has become sufficiently commonplace that some writers even attribute the phrase 'natural kind' to Mill; e.g. Alexander Bird and Emma Tobin write, "J. S. Mill... was one of the first to use the phrase 'natural kind' " [1]. Mill never used the phrase, however, even though his critics use the phrase consistently later in the 19th century.[1]

So the narrative involves two claims of continuity: first, that recent debates are continuations of ones that began with Mill; second, that the term of Mill's system which maps onto our 'natural kind' is his 'Kind'.

Both these claims are mistaken, but my focus here is on the second.[2] Mill's terms do not map one-to-one onto ours. In addition to Kinds, Mill has an account of *natural groups*. Mill's natural groups and Kinds answer two different questions about what we call natural kinds.[3]

The first question is about what, as a matter of taxonomy, distinguishes natural kinds from arbitrary categories: What criteria must a category satisfy to count as a natural kind? This is not particularly an epistemic matter, because we might not and perhaps could never be in a position to apply the criteria. However, it is metaphysically somewhat thin. An answer to it specifies what a category must do in order to fulfill the natural kind rôle, but it need not specify the fundamental ontology of such categories.

The second question concerns ontology: What kind of being has a natural kind got? Answers might appeal to causal structure, universals, or primitive similarity.

Call these the *taxonomy* and *ontology* question, respectively.[4]

The two questions are conflated in many recent discussions of natural kinds. If we answer the taxonomy question by saying that natural kinds are those which carve nature at its joints, then we answer the ontology question in terms

[1]It is unclear exactly when Mill's Kinds came to be called 'natural kinds' as a matter of jargon. Hacking [4] attributes the phrase 'natural kinds' to John Venn, and the attribution is part of the standard narrative; for example, it is repeated uncritically by Laura Snyder [15, p. 157, fn. 2]. Although Venn uses the words 'natural' and 'kind' together, it is unclear that Venn was responsible for 'natural kind' as a fixed phrase; cf. Magnus [9, pp. 2–3].

[2]Magnus [9] debunks the first claim of continuity, arguing that the recent vogue for *natural kinds* is not a continuation of 19th-century debates using the same phrase.

[3]Hawley and Bird [6] call these the 'naturalness' and 'kindhood' questions, respectively, and point out that the distinction is not typically made. I have argued for the importance of the distinction in the context of Homeostatic Property Cluster accounts [8].

[4]Even though the labels are mine, rather than Mill's, it is clear that natural groups and Kinds play two different rôles in his system. So (I argue) it is not anachronistic to see them as answers to different questions.

of nature's joints. We discharge both questions at once, and so it would be gratuitous to distinguish them. The same elision occurs in more sophisticated accounts. For David Lewis and followers, natural properties are "an élite minority of special properties" [7, p. 346] and that eliteness is a matter of fundamental metaphysics. A category is a natural kind if and only if it corresponds to a natural property, providing taxonomy and ontology altogether. Similar elision follows for any essentialist account in which natural kinds stand in a one-to-one relationship with essences.

My central claim here is that Mill gives the two questions importantly different answers — as a matter of history, Mill's categories cannot be neatly mapped onto contemporary terms. At the end, I briefly suggest how we might profit by minding the distinction that Mill made in the 19th century but which was lost in the 20th.

2 Whewell

This section briefly considers some features of William Whewell's account of classification. As we will see, Mill explicitly engages Whewell, and the contrast between their views highlights Mill's innovation.[5]

Whewell claims that the aim of taxonomy is to provide a natural classification, to divide things into *kinds* or — as he more often writes — *natural classes*. These are the categories that will support systematic induction. He writes that "since the truths we are to attend to are scientific truths, governed by precise and homogeneous relations, we must not found our scientific Classification on casual, indefinite, and unconnected considerations" [18, p. 115].

Importantly, for Whewell, natural classes will support scientific inference because they reflect the underlying construction of the world. So taxonomy aims not merely to organize things for science but also to discover the world's construction. Discussing mineralogy, Whewell writes, "the science which we require is a complete and consistent classified system of all inorganic bodies. For chemistry proceeds upon the principle that the constitution of a body invariably determines its properties; and consequently, its kind. . . " [17, p. 189]. Discussing botany, he writes similarly,

> No person, however, who wishes to know botany as a science, that is, as a body of general truths, can be content with making names his ultimate object. Such a person will be constantly and irresistibly led on to attempt to catch sight of the natural arrangement of plants,

[5] Mill explicitly acknowledges Whewell as providing him the crucial clue to Kinds. He had stopped working on the *Logic* for five years, because he was unable to make sense of induction. But Whewells' 1837 *History of the Inductive Sciences* provided him with the comprehensive survey of physical science which he needed to move ahead. [11] Although Mill found much to disagree with in Whewell's philosophy, there are considerable similarities in their accounts of natural classification. Mill quotes Whewell approvingly on the topic [12, p. 488] and, where he disagrees, still quotes Whewell at some length [12, p. 501–2]. For more on the relation between Whewell and Mill on classification, see McOuat [10] and Snyder [15].

> even before he discovers, as he will discover by pursuing such a course of study, that the knowledge of the natural arrangement is the knowledge of the essential construction and vital mechanism of plants. [17, pp. 319–320]

So what makes kinds natural for Whewell is ultimately the "constitution" and "construction" of things. The taxonomy and ontology questions are answered together.

3 Mill

Initially in Book I of the *Logic*, Mill distinguishes Kinds (with a capital-K) from arbitrary classes. A class can be indicated by any property or list of properties. For example, the class of *white* things corresponds to the property of being white, and the class of *red round* things corresponds to the properties of being red and of being round. Because there is a class corresponding to any property or list of properties, no such class is more natural than any other. White things have nothing in common beyond their whiteness and its necessary consequences (e.g., that all white things are non-transparent). In contrast, Kinds are classes of things which share indefinitely many properties. There are some diagnostic criteria which we associate with a chemical kind or biological species, but the members share many properties besides those which we use to mark the Kind. On Mill's view, a Kind "is distinguished from all other classes by an indeterminate multitude of properties not derivable from one another" [12, p. 99].

For Mill, Kinds are crucial for inductive generalization. Suppose we subject a sample of phosphorus to an experimental condition in the lab and we infer that other samples of phosphorus will react similarly. This relies on the other phosphorus, the stuff outside the lab, sharing enough properties with our sample that the condition happening to them counts as an instance of the same cause. We identify other samples of phosphorus merely by diagnostic criteria, so how can we rely on distinct bits of phosphorus sharing further properties beyond those used to diagnose them as phosphorus? We can do so, Mill would say, because phosphorus is a Kind. The diagnostic criteria identify samples as members of the Kind, assuring that they share indefinitely many other features.

In this example, the fact that all lumps of phosphorus are the same Kind is crucial to a causal inference about what things like this will do. Yet, because of Mill's conception of causation, Kinds cannot themselves be held together by causes. Mill thinks of causal inference as guided by the *law of causation* which states that every event is preceded by some circumstances which necessitate it: When those circumstances occur, the effect invariably follows [12, p. 410]. This means that causes are regularities that obtain between prior and subsequent events.

Kinds are also regularities, but they obtain between different things at the same time (e.g. all the samples of phosphorus) rather than between events at different times (e.g. heatings of phosphorus and a subsequent ignitions). For

a Kind, Mill writes, the the shared properties are an "invariable co-existent, in the same manner as an event must have an invariable antecedent" [12, bk. III, ch. XXII]. Kinds are structures of non-causal regularities.

Laura Snyder describes Mill as "*denying* that kinds are natural" and writes that Mill's Kinds "are not real kinds" [15, p. 164]. What she means by this is that, for Mill, there is no underlying mechanism "causally responsible for the production of... shared superficial qualities" [15, p. 164]. She is correct that Mill's Kinds do not have a real essence in Locke's sense, that there is no deeper and more fundamental process which causally produces the regularity observed in members of the Kind. Unlike Whewell, Mill refuses to talk about the constitution or essential construction of things. However, Mill's Kinds are not enquiry dependent or merely nominal.[6] Each corresponds to a law of nature, a law of coexistence which has the same reality as diachronic causal laws. They are defined in terms of the course of events, rather than in terms of actual or possible science.

In Book IV, Mill takes up "operations subsidiary to induction" such as observation, abstraction, naming, and classification. In discussing naming, Mill explicitly invokes the conception of Kinds which he developed in Book I.[7] In discussing classification, Mill makes a different distinction between natural groups and merely technical or artificial ones. He says some natural groups will be Kinds but that not all of them will be. Natural groups — in contrast to Kinds — are characterized by their function in scientific enquiry.

Properly scientific classification, in order to be as general as possible, should reflect the causal structure of things. It is best "when the objects are formed into groups respecting which a greater number of general propositions can be made.... The properties, therefore, according to which objects are classified should, if possible, be those which are causes of many other properties..." [12, p. 499]. He distinguishes properly natural classification from artificial classification; continuing, "A classification thus formed is properly scientific or philosophical, and is commonly called a Natural, in contradistinction to Technical or Artificial, classification or arrangement" [12, p. 499]. The categories that figure in a natural classification he calls *natural groups*. Mill uses the adjective 'natural' here to discuss natural classification and natural groups, but he never uses it to modify Kinds. The phrase 'natural kind' was not part of his vocabulary.

[6]Mill writes that "there are in nature distinctions of Kind; distinctions not consisting in a given number of definite properties *plus* the effects which follow from those properties, but running through the whole nature... of the things so distinguished" [12, p. 502].

[7]Mill calls a system of names for Kinds 'nomenclature', in contrast to mere 'terminology'. Lavoisier's new chemistry and Linnæus' system of biology, he writes, provided nomenclature. The taxonomic innovations allowed enquiry to move beyond parochial concerns, to chart Kinds rather than mere categories of interest. Having a nomenclature is the mark scientific progress, Mill thinks, and in other fields a lack of nomenclature "is now the principle cause which retards the progress of the science" [12, p. 492]. Mill defines 'nomenclature' explicitly by reference to Kinds, as "the collection of names of all the Kinds with which any branch of knowledge is conversant"[12, p. 492]. He takes this distinction from Whewell. Mill writes, "The words Nomenclature and Terminology are employed by most authors almost indiscriminately; Dr. Whewell being, as far as I am aware, the first writer who has regularly assigned to the two words different meanings" [12, p. 492].

Mill insists that science will need names for more than just Kinds. He does think that Kinds will should appear in a proper scientific classification, and so Kinds qualify as natural groups — but he insists that a complete classification will require more categories than there are Kinds. He writes, "The distinctions between Kinds are not numerous enough to make up the whole of classification" [12, p. 503].

The natural groups which are not Kinds distinguish the important qualities of things. This is subject to the worry that *importance* depends on human concerns. Mill recognizes this worry, acknowledging that farmers divide plants differently than botanists and that geologists divide fossils differently than zoologists [12, p. 500]. If this were the end of it, then natural groups (besides those which correspond to Kinds) would not be real features of the world. They would be determined by our sense of what is important, shaped by our projects and interest. Different concerns could make for different taxa.

Mill avoids this result by saying that the natural groups are the ones which would figure in the science of a disinterested enquirer. He writes that

> when we are studying objects not for any special practical end, but for the sake of extending our knowledge of the whole of their properties and relations, we must consider as the most important attributes those which... would most impress the attention of a spectator who knew all their properties by was not especially interested in any. Classes formed on this principle may be called, in a more emphatic manner than any others, natural groups. [12, pp. 500–1]

Natural groups would be identified by an ideal, neutral observer. So they are objective in the sense of not being dependent on any particular subjective standpoint.

Mill's characterization of natural groups as the categories of an intersubjectively warranted taxonomy diverges from his characterization of Kinds as determined by objective laws of coexistence. The two characterizations do not pick out the same categories, and their rationale is importantly different. Natural groups are defined in terms of possible or ideal enquiry, whereas Kinds are defined just in terms of how the world is.

By contrast, although Whewell provides characterizations of Natural Classes both as objects of possible enquiry and as features of the world, for him the difference is just one of exposition. As we saw, Whewell thought that ideal scientific enquiry should divide things by their essential constitutions.

To put the difference in our terms, we might approach natural kinds by way of taxonomy or by way of ontology. For Whewell, this makes no difference, and any legitimate scientific categories can be approached from either direction. For Mill, the two do not perfectly coincide. Beginning with taxonomy, we get a wealth of natural groups. Beginning with ontology, we get just the Kinds.

4 Mill's critics

In an 1887 attack on Mill's "doctrine of natural kinds", M.H. Towry enumerates four objections to Mill's account.[8] For our purposes, we can treat them as raising two broader worries.[9]

One worry is epistemic and semantic. According to Mill, we frame an arbitrary class by stipulating properties which hold of its members. The class of *white things* is specified just by the property *white*. Towry accepts this and argues that the same is true for all classes and kinds. She writes:

> Nature has in reality neither the class White Things nor the class Horse. We made both. ... There are a quantity of things in the universe, alike in point of being white; there are a quantity of things alike in points *a b c*, &c. = Horses. The properties are not found by the Kind, but the Kinds are formed by the properties. [16, p. 436]

So, Towry writes, "one class is no whit less a merely intellectual creation than the other" [16, p. 436].

Another worry is metaphysical. Mill posits a difference in kind between Kinds and mere classes, but Towry objects that there is at most a difference in degree. There are anomalies and intermediate cases. Towry invokes Whewell, writing that "Whewell's type-theory seems to me nearer the truth than Mill's impassable barriers, because it recognizes infinite gradations and interminglings" [16, p. 438]. But Towry dissents from both Whewell and Mill by insisting that Kinds are just nominal classes. She writes, "When we advance beyond Singulars to many individuals or substances forming a 'natural Kind,' we have made an arbitrary and conventional combination" [16, p. 438]. That is to say, the Kind does not correspond to anything in nature.

I think that Mill can fairly be seen to anticipate the first worry. He recognizes that the semantics for Kinds must be different than the semantics for stipulated groups, and so he holds that the term for a Kind has a different connotation than the term for an arbitrary class. The term for an arbitrary class consists merely of some stipulated attributes. The term for a Kind consists of some attributes which distinguish the class along with the commitment to that class's being a Kind.[10]

[8] Although Franklin and Franklin (whom I discuss below) address their reply to "Mr. Towry", it seems likely that the author was Mary Helen Towry White. She published on a range of topics — from the history of Scottish clans to stories of famous children — and was credited under different variations of her name. My inability to decisively confirm that this is the same Towry is an example of how women who contributed to philosophical debates are made to disappear from our retelling of them.

[9] Towry begins with a fair and concise summary of Mill's view: "Mill says that a Kind is one of those classes which are distinguished from all others, not by one or a few definite properties, but by an unknown multitude of them; the combination of properties on which the class is grounded being a mere index to an indefinite number of other distinctive attributes, and instances Plant, Animal, Sulphur, Horse, &c., as Kinds" [16, p. 435].

[10] Regarding terms for Kinds, Mill writes, "besides connoting certain attributes, they also connote that those attributes are distinctive of a Kind" [12, p. 493]. This is an explicit point of contrast with Whewell. On Whewell's account, we identify an exemplary individual as the

I think that Mill also has a ready response to Towry's second worry, because he only introduces Kinds as a way to understand how inductive generalization is possible. If Towry's worry were legitimate, then there would be no difference in the world between real groups (like *phosphorus*) and an arbitrarily concatenated group (like the union of *phosphorus* and *sandwiches*) — but then there would be no more ground to generalize from samples of phosphorus than from samples of phosphorus-or-sandwiches. This point is especially clear in hindsight, because we are familiar with Goodman's new riddle of induction. Even though philosophers may disagree about what distinguishes 'emerald' from 'emerose' (where 'emerose' picks out all the observed emeralds and all the unobserved roses) it is clear that something does. To put the point in terms which were available to Mill's 19th-century critics: Making sense of science requires that there be some distinction between arbitrary and non-arbitrary classes. Insofar as Mill is aiming at that distinction, there is something right about his notion of Kinds.[11]

There are two published replies to Towry.[12] In the second of these, Fabian Franklin and Christine Ladd Franklin concede to Towry that there may be no fundamental difference between the mental operations by which we come to think of arbitrary classes and natural kinds but insist that there is nonetheless an important difference between them in the world. They begin, "The doctrine of Kinds, as laid down by Mill, does not seem to be tenable... yet there is, we think, a real difference between such classes and mere arbitrary classes; and the nature of that difference may be stated very nearly as Mill stated it" [3, p. 83]. Although they accept that any category is "an intellectual creation" they maintain that it could not be "a *merely* intellectual creation" [3, p. 84].

Mill's mistake, the Franklins suggest, was to suppose that what holds a Kind together is a fundamental non-causal regularity which cannot be explained. Rather, they suggest that the connection can be explained by either a causal regularity or a historical connection between different members of the Kind. They write:

> When a certain set of qualities entails the presence of others, and the supposition cannot be entertained that there is a causal connexion of a general nature between them, the conclusion is inevitable... that

type, and the Kind is the class of things which are sufficiently similar to the type specimen. On Mill's account, we identify a list of properties which are diagnostic of the Kind, and the Kind is the class of things which share the diagnostic properties and indefinitely many more. As Whewell would have it, we read the diagnostic properties off of a designated type specimen. Mill allows that we can imagine a type specimen, but he thinks that we do so by imagining a thing with all of the diagnostic properties. [12, pp. 501–5] Schwartz [14] provides an extended discussion of Mill's semantics for Kind terms.

[11] One might worry that my reading of Mill describes Kinds as independent of enquiry, but the reply to Towry defends Kinds by appealing to the possibility of enquiry. Such a worry is easily defused: Although making sense of enquiry provides Mill's reason for positing Kinds, Kinds are not defined in terms of enquiry.

[12] In the first of these, W.H.S. Monck [13] insists that taxonomy is not a subject which should be addressed by a logician at all, since it concerns knowledge of what the world is actually like. This objection is oddly hidebound. It is obvious in the sections on Kinds and categories that Mill, like Whewell before him, is doing philosophy of science.

> there is *a certain community of origin* among the objects possessing that set of qualities. [3, p. 84]

By 'community of origin' they mean some common cause; that is, that members of a natural kind have a shared history which explains their shared features.

Common cause provides a way to explain regularity, without it being the unconditional result of causal or non-causal laws. Because of their common history, the members of such a Kind will share features beyond those which we initially notice or by which we diagnose membership in the Kind; when "we regard the invariable concomitance of certain qualities with certain other marks as proof of a common origin in the objects possessing those marks, there is no reason for setting any limit to the number of ways in which that common origin will be betrayed" [3, p. 85].

A consequence of this proposal is that Mill's exemplary Kinds turn out to be a disparate lot. Biological species are groups of common descent, and so they can be explained by community of origin. Yet chemical kinds do not seem to be. Rather, it seems more likely that chemical regularities are a matter of causal law. Considering the example of sodium, the Franklins write, "there is not... any external evidence that all the sodium... in the universe was derived from a common stock; but it seems highly probably that either this is the case or else that all the properties of sodium are deducible by general laws from a few of them... [that] the properties of sodium are deductions from its molecular constitution" [3, p. 85]. The only general thing to say about Kinds is that their unity can be explained *either* by general laws *or* by common causes — i.e., "either the qualities or the objects have a real connexion with each other" [3, p. 85].

This furthers the division between taxonomy and ontology that we saw already in Mill's account. For Mill, some but not all natural groups correspond to Kinds in the world. So the characterization of the criteria for what makes a category natural is separate from the metaphysical description of what it is in the world which satisfies those criteria. Franklin and Franklin drive the wedge further, by suggesting that different categories might be realized in the world in fundamentally different ways. Some natural groups, like chemical elements, are unified because members of the kind have a similar composition and so behave similarly according to general, causal laws. Others, like biological species, are unified by sharing a historical source and so behave similarly because of their common cause.

5 Conclusion

If we treat the 19th-century discussions as an anticipation of debates about natural kinds in the last 50 years, Mill has two separate notions which might be mapped onto our present term 'natural kind': Kinds and natural groups. As is usual in the history of philosophy, it would be a gross over-simplification to treat this simply as a matter of translation. The fact that there is not one clear

counterpart to our term 'natural kind' suggests that, in some sense, Mill was not thinking about *natural kinds* the way that we do.

We should not pretend that Mill had two entangled notions where we now simply have one. Quite the contrary, we can distinguish the taxonomy and ontology questions about what we call *natural kinds*. First, what criteria distinguish natural kinds from arbitrary categories? Second, what features of the world make some categories but not others satisfy these criteria?

Failure to mind this distinction can be seen to lead to confusion in recent debates. Establishing this is beyond the scope of this paper, but I will point to one suggestive illustration: The idea that natural kinds are homeostatic property clusters (HPCs) is most plausible if we treat it as an answer to the ontology question for many but not all natural kinds. Yet many authors respond to HPC accounts just by providing examples of natural kinds which are not HPCs or of HPCs which are not natural kinds. Those counterexamples are only relevant if we take HPCs as an answer to the both questions, to define both what it is to be a natural kind and what a natural kind is in the world.[13]

We should reject the usual historical account, according to which Mill's Kinds matured into our natural kinds. We understand Mill better if we recognize that he was struggling with separate issues, and that he introduced several notions to resolve them. To revisit Hacking's metaphor: The scholastic darkness which shadows present discussions of natural kinds may be dissolved not by abandoning natural kinds altogether but by recognizing complexities too often overlooked. We would do well to let a Millian flower bloom.

References

[1] Alexander Bird and Emma Tobin. Natural kinds. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2009. Available from: `http://plato.stanford.edu/archives/spr2009/entries/natural-kinds/`.

[2] John Dupré. Foreword. In Joseph Keim Campell, Michael O'Rourke, and Matthew H. Slater, editors, *Carving nature at its joints: Natural kinds in metaphysics and science*, pages vii–viii. The MIT Press, Cambridge, Massachusetts, 2011.

[3] F[abian] Franklin and C[hristine] L[add] Franklin. Mill's natural kinds. *Mind*, 13(49):83–85, January 1888.

[4] Ian Hacking. A tradition of natural kinds. *Philosophical Studies*, 61:109–126, 1991.

[5] Ian Hacking. Natural kinds: Rosy dawn, scholastic twilight. *Royal Institute of Philosophy Supplement*, 82:203–239, 2007.

[13] I discuss this example at greater length elsewhere [8].

[6] Katherine Hawley and Alexander Bird. What are natural kinds? *Philosophical Perspectives*, 25(1):205–221, December 2011.

[7] David Lewis. New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377, December 1983. doi:10.1080/00048408312341131.

[8] P.D. Magnus. NK≠HPC. *Philosophical Quarterly*, 64(256):471–477, July 2014. doi:10.1093/pq/pqu010.

[9] P.D. Magnus. No grist for Mill on natural kinds. *Journal for the History of Analytical Philosophy*, 2(4):1–15, 2014.

[10] Gordon McOuat. The origins of 'natural kinds': Keeping 'essentialism' at bay in the age of reform. *Intellectual History Review*, 19(2):211–230, 2009. doi:10.1080/17496970902981694.

[11] John Stuart Mill. *Autobiography*. Project Gutenberg, [1873] 2003. Available from: http://www.gutenberg.org/ebooks/10378.

[12] John Stuart Mill. *A System of Logic*. Harper&Brothers, New York, eighth edition, 1874.

[13] W. H. S. Monck. Mill's doctrine of natural kinds. *Mind*, 12(48):637–640, October 1887.

[14] Stephen P. Schwartz. Mill and Kripke on proper names and natural kind terms. *The British Journal for the History of Philosophy*, forthcoming.

[15] Laura J. Snyder. *Reforming Philosophy*. University of Chicago Press, Chicago, 2006.

[16] M. H. Towry. On the doctrine of natural kinds. *Mind*, 12(47):434–438, July 1887.

[17] William Whewell. *The History of the Inductive Sciences*, volume III. John W. Parker and Son, London, 1837.

[18] William Whewell. *The History of Scientific Ideas*, volume II. John W. Parker and Son, London, third edition, 1858.

EXPERIMENT AND ANIMAL MINDS: WHY STATISTICAL CHOICES MATTER

PSA 2014
Forthcoming in *Philosophy of Science*

Irina Mikhalevich, Washington University- St. Louis

ABSTRACT

Comparative cognition is the interdisciplinary study of nonhuman animal cognition. It has been criticized for systematically underattributing sophisticated cognition to nonhuman animals, a problem that I refer to as the *underattribution bias*. In this paper, I show that philosophical treatments of this bias at the experimental level have emphasized one feature of the experimental-statistical methodology (the preferential guarding against false positives over false negatives) at the expense of neglecting another feature (the default, or *null*, hypothesis). In order to eliminate this bias, I propose a reformulation of the standard statistical framework in comparative cognition. My proposal identifies and removes a problematic reliance on the value of parsimony in the calibration of the null hypothesis, replacing it with relevant empirical and theoretical information. In so doing, I illustrate how epistemic and non-epistemic values can covertly enter scientific methodology through features of statistical models, potentially biasing the products of scientific research. Broadly construed, this paper calls for increased philosophical attention to the experimental methodology and statistical choices.

INTRODUCTION

Comparative cognition is the interdisciplinary study of the evolution, development, and function of cognitive processes and mechanisms in nonhuman animals. A central controversy within the field concerns ways of guarding against bias in the course of interpreting nonhuman animal (henceforth, animal) behavior. As many philosophers and scientists have written, the worry is disproportionally aimed at guarding against overattribution of sophisticated cognition to animals (Sober 2005, Andrews 2011, de Waal 1998). For the purposes of this paper, I take it as a given that comparative cognition researchers as a group prefer explanations with the most austere cognitive ontologies, and that this practice results in an *underattribution bias*, or the systematic underascription of putatively complex cognition to animals (Andrews 2011, Fitzpatrick 2008, Meketa 2014). This preference is typically cashed out in terms of taking putatively simple mechanisms, processes, or abilities as the default experimental hypothesis.

In this essay, I locate the mechanism that drives the underattribution bias within the choice of the statistical *null* hypothesis (H_0). I argue that the manner in which the null hypothesis is currently chosen embeds a parsimony-based preference for simple cognitive ontologies. Having identified the mechanism driving the underattribution bias, I recommend removing that mechanism from the statistical methodology in which it is embedded, and replacing it with a procedure that is sensitive to empirical information. In so doing, I offer a case study of how values, such as parsimony, may come to play a central, though implicit, role in scientific methodology.

1. LOCATING THE SOURCE OF THE UNDERATTRIBUTION BIAS: THE NEYMAN-PEARSON METHOD OF HYPOTHESIS-TESTING

In the case of comparative cognition, the dominant statistical analysis method is what is known as the Neyman-Pearson Method (NPM) of hypothesis testing. Although the NPM is not the only statistical system available to science – there are also Bayesian and likelihoodist methods – it is the orthodoxy in comparative psychology, and, by extension, in comparative cognition.[1] The NPM includes what Peter Godfrey-Smith (1994) calls the *error-rate asymmetry*, which calls for preferring one type of hypothesis over another. Because the hypothesis typically preferred in comparative cognition is the one positing the simplest cognitive ontology, the error-rate asymmetry results in the underattribution bias. The remainder of this section explores the results of the error-rate asymmetry and sets the stage for my proposed solution for eliminating the underattribution bias. My solution works within the dominant paradigm of the NPM, retaining its desirable features, but offering a means of eliminating the parsimony-based underattribution bias. Put another way, my solution should be viewed as a *reformation* rather than as a *revolution*.

What exactly is the NPM? Put simply, the NPM is a method for controlling the error-rates (long-run relative frequencies) of two types of errors, which are labeled Type I errors and Type II errors. Type I errors, in general, are *defined* as those that are most serious. In the Neyman-Pearson tradition, the assumption is that the most serious type of error is the one that rejects the null hypothesis (H_0) when the H_0 is true. Within this paradigm, accepting the H_0 when

[1] For challenges to the orthodoxy of the NPM, see Kruschke (2010), who advocates replacing it with Bayesian analysis, and Anderson et al. (2000), who favor a version of likelihoodist methods to the allegedly "unscientific" null hypothesis testing methods. For philosophical defenses of the NPM, see Mayo (1981) and Mayo (1992). For a "severity-analysis" reformulation of the NPM see Mayo (2004) and Mayo and Spanos (2009, 2011).

it is false is a Type II error, and it is treated as less serious. Type I error *rates* are denoted as α and Type II error rates as β. To modify slightly Dienes's (2008) formalization of the relationship between error types and their relative frequencies, we may say that

$$\alpha =_{def} \mathbf{P(rejecting\ H_0 \mid H_0)}$$
$$\beta =_{def} \mathbf{P(accepting\ H_0 \mid \neg H_0).}$$

Although the NPM provides a means for controlling error rates in a way that minimizes the risk of making both types of errors, researchers have traditionally set the risk of a Type I errors lower than Type II errors. Typical values for α are .05, .01, and sometimes .001. Treating Type I errors as more serious translates into controlling for Type I errors by making α very small, while keeping β either large or not controlled at all. This preference for making Type II errors over Type I errors is the *error-rate asymmetry*.

2. THE IMPORTANCE OF THE NULL HYPOTHESIS

With this background complete, it is now possible to see how earlier assessments of the underattribution bias in comparative cognition have correctly located the source of the bias at the statistical level, but preferentially emphasized the error-rate asymmetry over what I will now argue is the real mechanism: the *null hypothesis*.

Earlier Solutions: Locating the Underattribution Bias in the Error-Rate Asymmetry

Elliott Sober (2001, 2005) argues that there is no reason to prefer making Type II errors over Type I errors in comparative cognition and that this preference is furthermore a misapplication of MC, understood as a parsimony principle. According to Sober, not only is MC *not* a parsimony principle, but both types of errors are equally undesirable, since both errors are

equally wrong. Sober advocates ridding the field entirely of the error-rate asymmetry, arguing that, "the only prophylactic [against risk of error] is empiricism" (Sober 2005, 97).

Building on Sober's work, Andrews (2011) identifies the preference for Type I error with an exaggerated and damaging worry over the alleged systematic overattribution of sophisticated cognition to animals.[2] She agrees with Sober that the matter is an empirical one, but departs from his conclusion regarding the seriousness of each type of error. She argues that Type I errors are in fact *more* damaging than Type II errors because they foreclose on the possibility of future research (Andrews 2011). On her view, preferring to make Type I errors means preferring to wrongly conclude that, e.g., the New Caledonian crows do not use planning to solve puzzles. Once such a judgment has been made, it no longer makes sense to ask further questions about the features of e.g., the crows' future-planning abilities, such as whether they are domain-specific or general, available only with appropriate environmental scaffolding, and so on. As a result, a potentially fruitful research program never gets a chance to get off the ground. Type II errors, on the other hand, promote a further refinement of experimental questions. These questions may produce results that conflict with the original (mistaken) judgment, but, argues Andrews, science must be willing to take such risks.

Moving to the Null Hypothesis

Despite discussing the biasing effects of the error-rate asymmetry, neither Sober (2001, 2005) nor Andrews (2011) question the fact that the H_0 is treated as the absence of the mental feature

[2] Andrews refers to this overattribution fear as the fear of so-called anthropomorphism, or the attribution of allegedly uniquely human cognitive complexity to nonhuman animals. For in-depth analyses of the alleged mistake of anthropomorphism, or the mistaken attribution of human properties to nonhuman entities, in comparative psychology, see Fisher (1990; 1991). See also Keeley (2004) for an update to Fisher's arguments.

under investigation.[3] However, as I will now show, the choice of the H_0 is just as likely to be a source of bias as the error-rate asymmetry. Once the H_0 is understood to be a source of bias, a solution to the underattribution problem will become clear.

Consider Andrews's claim that the error-rate asymmetry results in what she calls a behavioristic bias. She is right as long as the H_0 is defined as the absence of a cognitive feature. However, if the H_0 were defined as the *presence* of rich cognitive abilities, the result would be the *opposite* of a bias toward underattribution: comparative cognition would be biased toward *overattribution*. Such a dramatic difference in the outcome of the application of our procedural rules points to the significance of the construction of the H_0, i.e., the choice of how it is to be defined. If the construction of the H_0 is so important to the final outcome of a given hypothesis-testing procedure, then we must pay more careful attention to how we come to identify something as the H_0.

To illustrate the importance of attending to the construction of the H_0 more concretely, consider a case where replacing the H_0 while retaining the error-rate asymmetry results in a bias toward sophisticated cognitive explanations. Let us take a closer look at an experiment by Allison Foote and Jonathom Crystal (2007), which used a duration-discrimination task to test for metacognition – awareness of one's own mental states – among rats. In this case, the mental state in question was that of uncertainty. Meketa (2014) describes the experiment as follows:

> "[Rats] were presented with audio tones of different durations and trained to classify the tones into the categories of "short" or "long." The rats were then presented with a range of tones, some clearly short and others ambiguous. Correct responses were rewarded with food, and incorrect responses were not rewarded at all. Next, the rats were given

[3] Since the present essay was written, Andrews and Brian Huss have written, but, to my knowledge, not yet published a manuscript that includes an explicit discussion of the role of the null hypothesis (Andrews & Huss *unpublished manuscript*).

> the same test, but were given a third option: to decline a test. Declined tests allowed the subjects to move on to more tests with the prospect of getting more food. When given the choice to decline tests, the rats consistently opted to decline the ambiguous ("harder") tests but not the unambiguous tests, even though declining a test resulted in a smaller food reward than answering correctly. Moreover, the overall accuracy improved when rats were allowed to opt out of difficult tests. Foote and Crystal (2007) concluded that the rats were aware of their own uncertainty." (Meketa 2014)

In other words, Crystal and Foote concluded that this behavior showed that the rats were metacognitive.

Let us now abstract away from the details of the experimental setup and just consider the hypothesis being tested**.** We see that their H_0 was that the rats do *not* possess metacognition. The alternative hypothesis – the one they wished to demonstrate – was that the rats *are* capable of metacognition. Given the error-rate asymmetry, the burden of proof falls on the metacognitive hypothesis. In fact, in a follow-up paper on metacognition, Crystal and Foote (2009) clearly state that the default hypothesis – the H_0 – is *and should be* that rats lack metacognitive capacities. The reason, they argue, is that the behaviors they observed in the 2007 trials *could* be explained by allegedly simpler mechanisms, such as associative learning, which is presumed to be incompatible with metacognition.[4] This means that the metacognitive explanations bear the burden of proof.

But now consider what would happen if the H_0 in Foote and Crystal's experiments were a rich cognitive explanation of the rats' behavior (e.g., H_0 = "rats are capable of metacognition").

[4] The standard view that associative mechanisms are different from and simpler than cognitive processes has been coming under scrutiny in recent years. For example, Cameron Buckner (2012*)* argues against the view that cognitive and associative systems are incompatible. Taking issue with the assumption that association is simple, Gallistel (2008) argues that associative mechanisms are more demanding than cognitive systems insofar as they would require far greater energy expenditures than alternative mechanisms. He uses the honeybee navigation system to argue that the honeybee brain does not have enough computing power to process information through associations alone, and must require a representational system of mental maps (Gallistel 2008).

Then the burden of proof would be on the hypotheses positing less sophisticated cognition (e.g., H_1 = "rats are relying on stimulus-response learning"). In this case, a preference for Type I errors over Type II errors would mean a preference for accepting (or failing to reject) the hypothesis that rats possess metacognitive abilities when, in fact, the rats do not. As a result, the underattribution bias would be inverted.

The metacognition example suggests that the way that the H_0 is constructed is *at least as important* as the error-rate asymmetry when it comes to assessing an experimental methodology for built-in theoretical commitments. What attending to the construction of the H_0 reveals is that, while the asymmetry introduces a bias, the nature of this bias is specified by the content of the H_0. In one sense, the role of the H_0 may be *more* important than the error-rate asymmetry: while the asymmetry can only be made more or less pronounced, the content of the H_0 can embed any number of problematic assumptions.

This conclusion prompts the question: Why, if the content of H_0 is so important, have scientists and philosophers of science assumed that the H_0 is naturally defined as "non-presence" or "no effect"? It is curious that a feature that carries such powerful implications for inference from experiment should be casually assumed to be globally fixed at the non-presence of the target cognitive property. In order to explain why the content of the H_0 has been systematically overlooked, I turn to Peter Godfrey-Smith's (1994) analysis of the NPM and the possible justifications for its use. Placing the NPM into its historical context will, furthermore, motivate my suggestion that the NPM can be modified as I suggest in §5.

3. THE NPM: A CHANGING JUSTIFICATION AND THE INTRODUCTION OF PARSIMONY

According to Godfrey-Smith (1994), the original justification for the NPM was pragmatic, but that justification was rejected shortly after its introduction while the method of preferentially

controlling for Type I errors was retained. Contrary to Andrews's claim, the original NPM included an accept/reject procedure. However, the original, pragmatic, justification of the 'accept/reject' decision-procedure was intended as a *behavioral* strategy, where "accepting" an hypothesis meant *acting as if* the hypothesis were true (Godfrey-Smith 1994, 280-82). This pragmatic justification meant that the NPM could not be used to support belief in the *truth* of an hypothesis or even in the *probability* of the hypothesis being true. This pragmatic justification did not sit well with subsequent scientists and statisticians (e.g., R.A. Fisher), who wanted a statistical system to provide *evidence* for the truth or falsity of an hypothesis – something that the original NPM explicitly avoided (Dienes 2008, Anderson et al., 2000, Gigerenzer 2004). The result, according to Godfrey-Smith, was a proliferation of alternative justifications that have in turn altered the method in unexpected ways.

One alternative justification – which Godfrey-Smith labels the 'semantic' justification[5] – includes the concept of what he calls a 'natural null,' or H_n, which is typically defined as the hypothesis of no effect or no difference. On the semantic justification, the Type I error is a wrong rejection of the hypothesis of no effect, or no difference. Since Type I errors are considered more serious, the semantic justification advises erring on the side of concluding that no effect or difference was detected. Moreover, the accept/reject procedure is interpreted both behaviorally and epistemically (Godfrey-Smith 1994, 287). I wish to focus on what Godfrey-Smith calls the 'semantic' justification for the NPM, because this is the most popular justification in psychology and, hence, also in comparative psychology and comparative cognition. It is worth noting that the other two justifications that Godfrey-Smith discusses, the 'pragmatic' and the 'doxastic,' do not specify a value for the H_0, holding that "H_0 is 'true' if the world is in a state

[5] Godfrey-Smith labels justification as "semantic" because it specifies the semantic content of the null.

such that the action associated with H_0 is better than the alternative action" (Godfrey-Smith 1994, 281).[6]

Crucially, Godfrey-Smith identifies a curious metaphysical principle embedded in the semantic NPM: When combined with the error-rate asymmetry, the H_n results in a preference for *nothing* over *something*, that is an Occamist commitment to maximally simple ontologies and theories that favor such ontologies. Since the semantic justification is, according to Godfrey-Smith's analysis, the most common interpretation in science – sometimes combined with the doxastic justification – it follows that the sciences that use it encode a commitment to Occamist metaphysics. Godfrey-Smith argues that psychology uses the semantic justification almost exclusively, though this is sometimes combined with a doxastic justification. Since comparative cognition is to a large extent constituted by comparative psychology, it is no surprise that Occamism is present in comparative cognition's statistical methodology as well.

Finally, if indeed the content of the H_0 is at least as significant as the error-rate asymmetry for identifying bias in comparative cognition, then Sober and Andrews have been focusing on a feature that only becomes a problem under conditions in which the null hypothesis is biased. It is possible to retain the asymmetry found in the NPM without accepting the question-begging conservatism about animal minds. My analysis recommends that the justifications for the NPM be carefully re-examined to avoid smuggling in a priori theoretical commitments.

[6] By contrast with this behavioral "pragmatic" interpretation, the doxastic justification for the NPM is epistemic. It replaces the pragmatic component with the rule that "when an observation in the critical region [the set of values that would cause us to reject the hypothesis] occurs the researcher rejects H_0. But when an observation falls outside the critical region the researcher *merely suspends judgment*" (Godfrey-Smith 1994, 282; emphasis added).

It is evident that statistical methods are often considered value-neutral and, in that respect, objective. However, I have shown that values may be embedded in these statistical methods. In the case of comparative cognition, this value is Occamist parsimony, and it is located in the choice of the null hypothesis within the NPM. Moreover, as a statistical methodology comes to be used as a standard in a given field, the values embedded in the method fade from scientific consciousness. The result is that, while researchers and philosophers appreciate the potential for a gerrymandering of data by cherry-picking statistical analyses, values, such as parsimony, continue to operate in the background methodology itself, without being subject to direct scrutiny. Analyses such as the one I offer here, are, therefore, crucial for uncovering and assessing the effects of values even in such inconspicuous places as tools for statistical analysis of experimental data. So much for a partial account of the invisibility of the null in the philosophy and science of comparative cognition. The next question is how the underattribution bias may be corrected. Given that comparative cognition researchers are unlikely to abandon the NPM in the near future, I propose a reformation of the semantic NPM.

4. THE NPM REFORMED: REPLACING THE NATURAL NULL WITH A CONTEXTUAL NULL

I have argued that the underattribution bias is driven by a parsimony-based preference for purportedly simple cognitive ontologies, and that this practice is regimented in the preference for a H_n. I will now show how to alter the "semantic" NPM in a manner that eliminates this bias. My strategy involves modifying the semantic NPM to replace the natural null with what I call a "contextual null" (H_c), which reflects a broader epistemic context for the animals under investigation.

To begin, note that the semantic view does not require that the H_0 must always be that the feature under investigation is absent. It is, however, this particular definition that lends the

semantic view its bias toward Occamist parsimony. It follows that removing the Occamism from the semantic NPM requires removing the natural null. This is precisely what I now suggest: the natural null should be replaced with a contextual null. In contrast to the H_n, the H_c is defined against a suite of background information about the research subjects, such as ontogenetic and phylogenetic information against the background of developmental and evolutionary theories, information about species-typical and individual behavioral profiles, neuroanatomical homologies[7] and homoplasies,[8] ecological context, and information from earlier studies and observational data.[9] I call such a null "contextual" for two reasons: (1) it respects the differences among experimental settings and the organisms being studied, and (2) it does not presuppose either cognitive complexity or cognitive simplicity. It is sensitive to the changing conditions between experiments, both in terms of the kinds of questions that are asked and with respect to how much is known about the target system.

The evolutionary considerations that I propose to be taken into account in constructing the H_c include the species' phylogenetic proximity to species about whom more is known in order to gauge the likelihood of homologous cognitive structures and abilities. These considerations already enter into decisions about which species to study when searching for a given ability, but not into the decision to cast a given hypothesis as the presumptive null. For example, chimpanzees' close phylogenetic proximity to humans is a frequently cited reason to test them for the presence of human-like abilities, such as tool use and metacognition.

---

[7] A homology is "a similarity inherited from a common ancestor" (Sober 2005, 94).
[8] A homoplasy is "a similarity that is the result of two or more independent derivations of the trait" (ibid).
[9] Fitzpatrick (2008) draws a very similar conclusion about the need for background information in hypothesis testing. However, he does not frame his case in statistical terms. His account is intended to displace the parsimony-based reading of Morgan's canon, understood as a heuristic, with a principle he calls "Evidentialism."

The developmental considerations relevant to constructing the H_c include, inter alia, hypotheses regarding developmental constraints on the evolution of the relevant cognitive and behavioral traits and the effects of the environment on gene expression. Ecological context would include information about the test subjects' behavior in its natural habitat, such as whether it is a social or solitary animal, whether it hunts or stores its food, whether it uses tools, and so on. Once again, these considerations already drive the research projects, suggesting that researchers consider such information to be probability-conferring. Consider the following example of research that has been guided by both ecological and developmental considerations. Furlong et al. (2008) tested chimpanzees for their ability to use tools based on the knowledge that chimpanzees use tools under natural conditions (e.g., dipping sticks into ant mounds to catch ants; using leaves to scoop up water). Based on the negative results obtained by a previous study by Daniel Povinelli (2000), which concluded that chimpanzees lack the competency for flexible use of implements, Furlong et al. hypothesized that Povinelli's chimps were developmentally stunted as a result of being brought up under socially impoverished conditions. Furlong et al. tested chimpanzees with different socialization backgrounds, and found that the ability to manipulate tools in a flexible manner (i.e., one suggestive of causal understanding) was positively correlated with social histories. This example shows the value of social ecology and its effects on chimpanzee intellectual development. Building such considerations into the null hypothesis would ensure that crucial information is not left out of the experimental design.

Finally, the H_c should include information from previous studies and observational data. However, given the arguments of the foregoing sections, including the work of earlier experimental studies would require re-evaluating them for the presence of bias. This can be achieved by analyzing the choice of null hypothesis to ensure that unwarranted metaphysical

preferences have not been smuggled in and that relevant empirical and theoretical information was included.

The upshot is that my proposed H_c ensures that Type I errors will always be more epistemically serious because probability-conferring evidence is built into the H_0. Sometimes this method will produce a H_0 positing a simple cognitive ontology, but this will no longer be based on a blanket Occamist preference for simple ontologies, but on an empirically-informed expectation. This suggestion respects the intuition that default hypotheses ought to be those that we have the best reason to adopt. In the end, my account preserves the risk-controlling structure of the semantic account of the NPM while eliminating the questionable metaphysics and replacing it with empirical information.

CONCLUSION

I have argued that earlier assessments of bias in comparative cognition at the level of statistical data evaluation have ignored an important feature of the orthodox NPM –namely the null hypothesis. I have suggested that this lacuna may be attributed to a specific interpretation of the NPM – the semantic justification – which assumes that the null hypothesis must be universally set to a natural null of "no difference" or "no effect." I have suggested that if the semantic version of the NPM commits the researcher to a position supported only by a problematic Occamist metaphysics, then the semantic version needs to be modified. My proposed modification to the NMP maintains the error-rate asymmetry, but replaces the H_n with the H_c. This change respects the intuition that the burden of proof should be on the hypothesis that has the least empirical and theoretical evidence on its side.

At a more general level, the foregoing discussion provides a case study of how metaphysical assumptions, such as a preference for simple ontologies, can enter science at the

level of statistical model choice. The fact that such metaphysical assumptions can be grafted onto the standard statistical models of an entire field suggests the need for a more careful scrutiny of statistical models, as philosophers of statistics, such as Deborah Mayo have been arguing for years. I showed that the data processing instruments used to generate inferences may not be value-free. Whether a value-free statistical instrument is desirable is an open question, but recognizing its value-laden dimensions is a necessary step in evaluating the conclusions drawn from scientific experiments.

REFERENCES

Anderson, David R., Kenneth P Burnham and William L. Thompson. 2000. "Null Hypothesis Testing: Problems Prevalence, and an Alternative." *The Journal of Wildlife Management* 64: 912-23

Andrews, Kristin. 2011. "Beyond Anthropomorphism: Attributing Psychological Properties to Animals." In *The Oxford Handbook of Animal Ethics,* edited by Tom Beauchamp and R.G. Frey. Oxford: Oxford U.P.

Buckner, C. 2011. "Two Approaches to the Distinction between Cognition and 'Mere Association.'" *International Journal of Comparative Psychology* 24: 314 – 348

Crystal, Jonathon D. and Allison L. Foote. 2009. "Metacognition in animals." *Comparative Cognition & Behavior* 4: 1 – 16.

Dienes, Zoltan. 2008. Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. *Palgrave Macmillan*.

De Waal, Frans B. M. 1999. "Anthropomorphism and Anthropodenial." *Philosophical Topics* 27: 255 – 280

———. 1997. Foreword in *Anthropomorphism, Anecdotes, and Animals,* edited by Robert W. Mitchell, Nicolas S. Thompson, and H Lyn Miles, xii – xvii. Albany: SUNY Press.

Fitzpatrick, Simon. 2009. "The primate mindreading controversy: a case study in simplicity and methodology in animal psychology." In *The Philosophy of Animal Minds,* edited by Robert W. Lurz, 237 – 257. New York: Cambridge U.P.

———. 2008. "Doing Away with Morgan's Canon." *Mind & Language* 23: 224-246.

Foote, Allison L. & Jonathon D. Crystal. 2012. "'Play it Again': a new method for testing metacognition in animals" *Animal Cognition* 15: 187 – 199

———. 2007. "Metacognition in the rat." *Current Biology* 17: 551 – 555.

Furlong, E.E, K.J. Boose, and S. T. Boysen. 2008. "Raking it in: the impact of enculturation on chimpanzee tool use." *Animal Cognition* 11: 83 – 97.

Godfrey-Smith, Peter. 1994. "Of Nulls and Norms." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1: 280 – 290.

Heyes, Cecilia. 2012. "Simple Minds: A Qualified Defense of Associative Learning." *Philosophical Transactions of the Royal Society: Biological Sciences* 367: 2695–2703

Keeley, Brian L. 2004. "Anthropomorphism, Primatomorphism, Mammalomorphism: Understanding cross-species comparisons." *Biology & Philosophy* 19: 521 – 540.

Kruschke, J. K. 2010. "What to believe: Bayesian methods for data analysis." *Trends in Cognitive Sciences* 14: 293-300.

Le Pelley, M. E. 2012. "Metacognitive Monkeys or Associative Animals simple Reinforcement Learning Explains Uncertainty in Nonhuman Animals." *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Accessed January 16, 2012. doi: 10.1037.a0026478

———. 1886. "On the Study of Animal Intelligence" *Mind* 11: 174 – 185.

Mayo, Deborah G. 2004. "An Error-Statistical Philosophy of Evidence." in *The Nature of Scientific Evidence,* ed. Mark L. Taper and Subhash R. Lele, 79-118. Chicago: Chicago University Press.

——.1992. "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?" *Synthese* 90: 233-262.

——. 1981. "In Defense of the Neyman-Pearson Theory of Confidence Intervals." *Philosophy of Science* 48: 269-280.

Mayo, Deborah & Aris Spanos. 2011. "Error Statistics," In *Handbook of the Philosophy of Science*, *Volume 7: Philosophy of Statistics*. Edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster. 153-198. Philadelphia: Elsevier Inc.

Meketa, Irina. 2014. "A Critique of the Principle of Simplicity in Comparative Cognition" *Biology & Philosophy* doi: 10.1007/s10539-014-9429-z

Sober, Elliot. 2009. "Parsimony and Models of Animal Minds." In *The Philosophy of Animal Minds,* edited by Robert W. Lurz, 237 – 257. New York: Cambridge U.P.

———. 2005. "Comparative Psychology meets Evolutionary Biology: Morgan's Canon and Cladistic Parsimony." In *Thinking with Animals: New Perspectives on Anthropomorphism,* edited by Lorraine Datson and Gregg Mitman, 85 – 99. New York: Columbia U.P.

Wimpenny J.H, A.A.S. Weir, L. Clayton, C. Rutz, A. Kacelnik. 2009. "Cognitive Processes Associated with Sequential Tool Use in New Caledonian Crows." *PLoS ONE* 4: 6471.

Reference Models: Using Models to Turn Data into Evidence[1]

Teru Miyake
Nanyang Technological University, Singapore

Abstract:
Reference models of the earth's interior play an important role in the acquisition of knowledge about the earth's interior and the earth as a whole. Such models are used as a sort of standard reference against which data are compared. I argue that the use of reference models merits more attention than it has gotten so far in the literature on models, for it is an example of a method of doing science that has a long and significant history, and a study of reference models could increase our understanding of this methodology.

1. Introduction

Reference models of the earth's interior play an important role in the acquisition of knowledge about the earth's interior and the earth as a whole. Such models are used as a sort of standard reference against which data are compared. Deviations between the observations one would expect if the reference model were an accurate representation of the earth, and actual observations, are used to make inferences about the earth's interior. Perhaps the most widely used such model in geophysics, the Preliminary Reference Earth Model or PREM[2] (Dziewonski and Anderson 1981), was completed in 1981, and it has been utilized for the construction of many other models through the end of the century (Ritzwoller and Lavely 1995).

There is a recent, growing literature focusing on the use of models in science (e.g. Morgan and Morrison 1999, Wimsatt 2007, Weisberg 2013). The use of models in a manner similar to the way in which reference models are used in geophysics is described by Wimsatt (2007), but he mentions these uses merely in passing in his

[1] The research for this paper was funded by Nanyang Technological University SUG No. M4080821. I would like to thank the NTU Philosophy Division, especially Lina Jansson, as well as George Smith and Michael Friedman.
[2] See Smith 2007 for an account of the history of seismology leading up to the construction of PREM.

discussion of neutral models in biology. Weisberg (2013) has a much more comprehensive and systematic account of models and their uses in science, but he does not specifically mention a use of models in the manner I will describe in this paper. I will argue that the use of reference models merits more attention than it has gotten so far in the literature on models, for it is an example of a method of doing science that has a long and significant history, one which has recently been described by Smith (2002) and Harper (2011) as "turning theory into evidence", and a study of reference models could increase our understanding of this methodology.

The aim of this paper is to contribute to the literature on models by first locating reference models relative to the general taxonomy of models described by Weisberg, and comparing them to the use of neutral models in biology as described by Wimsatt. I will then examine some possible desiderata for the construction of reference models, and then end the paper with some considerations about the connection between reference models and "turning theory into evidence".

2. Models and Idealization

I will start with Weisberg's picture of models because it is the most ambitious recent attempt to give a comprehensive account of models and their use in science, and it appears likely itself to become a standard reference on models for philosophers of science. From the standpoint of Weisberg's picture, Earth reference models would best be construed as target-directed models that utilize Galilean idealization (Weisberg 2013, 74-112). Target-directed models are models for which the modeler has a specific target in mind. For earth reference models, the target is clearly the interior of the earth. In Weisberg's picture of models, there are three different ways in which models can be idealized: Galilean idealization, minimalist idealization, and

multiple-models idealization. Galilean idealization involves the simplification of models with the aim of making them more mathematically tractable. Minimalist idealization involves the construction of models that include only difference-making factors that are necessary for a phenomenon, with the aim of constructing an explanation of a given phenomenon. Multiple-models idealization involves building multiple incompatible models of a single phenomenon, usually in the study of highly complex phenomena.

As we shall see, earth reference models involve Galilean idealization, so I want to examine this notion in more depth. Weisberg's discussion of Galilean idealization (2013, 99) depends heavily on the description given in McMullin (1985). Typically, there is some phenomenon of interest, but it is too complicated to model faithfully, so an initial simplified model is created. Then, this simplified model is used to improve our understanding of the phenomenon, and the simplified model is gradually made more realistic in a process that McMullin calls "de-idealization". Weisberg takes the whole purpose of Galilean idealization to be to deal with intractability, and thus "advances in computational power and mathematical techniques should lead the Galilean idealizer to de-idealize" (99).

Weisberg does not give very detailed examples of this process of de-idealization, but McMullin does.[3] The most detailed example he gives is the Bohr model of the hydrogen atom (McMullin 1985, 260-261). The Bohr model, in which the electron is in a circular orbit around the proton, could be used to predict the energy levels of the electron, which could then be compared to spectroscopic

[3] McMullin makes distinctions of his own regarding idealization, such as that between formal and material idealization. The Bohr model of the atom is given as an example of formal idealization. McMullin's distinctions might well cross-cut Weisberg's distinctions, and I do not want to complicate the picture here, so I will refrain from any discussion of McMullin's distinctions.

observations of hydrogen. More specifically, a theoretical value for the Rydberg constant could be calculated, which could then be compared to empirical measurements of this constant. McMullin says that at least three idealizations were being made here: the neutron is at rest, the orbit of the electron is circular, and relativistic effects are left out. Later on, successive corrections were made to the model which, McMullin claims, resulted in a closer fit between the model and reality. McMullin describes this process as one where the model "serves as the basis for a continuing research program", one in which the model starts off as a tractable model that has significant departures from reality, and this model is gradually filled in with more and more details.

Here, I want to ask exactly *how* the initial model serves as a basis for this research program. There are two significantly different ways in which it could do that. The first way is for the model simply to provide a sort of skeleton upon which further and further new details are added. These details might come about through new observations, or through the development of new mathematical or computational techniques that overcome the intractability problems that led to the development of the initial simplified model, allowing such details to be filled in, where previously they could not. The second way is for the model *itself* to be used directly to produce the new observations from which the further details can be added. I will call the first kind of process *passive* de-idealization, while I will call the second kind *active* de-idealization. We will see that earth reference models are used for active de-idealization.

Exactly how does active de-idealization work? Although Wimsatt (2007) does not use my terminology, he describes an example of active de-idealization. One of the major points that Wimsatt makes is that false models can be used in many

different ways to learn true facts about complicated systems. He gives a list of twelve ways in which false models can be used to search for better models. I want to focus here on the first five such functions he gives for false models:

> 1. An oversimplified model may act as a starting point in a series of models of increasing complexity and realism.
> 2. A known incorrect but otherwise suggestive model may undercut the too ready acceptance of a preferred hypothesis by suggesting new alternative lines for the explanation of the phenomena.
> 3. An incorrect model may suggest new predictive tests or new refinements of an established model, or highlight specific features of it as particularly important.
> 4. An incomplete model may be used as a template, which captures larger or otherwise more obvious effects that can then be "factored out" to detect phenomena that would otherwise be masked or be too small to be seen.
> 5. A model that is incomplete may be used as a template for estimating the magnitudes of parameters that are not included in the model.
>
> (Wimsatt 2007, 104)

The first function is, of course, mentioned by both Weisberg and McMullin. It is a statement of the idea of Galilean idealization and the process of gradual de-idealization. In functions 2 and 3, a false model is used as a heuristic—it suggests "new alternative lines for the explanation of phenomena", or "new predictive tests or new refinements". I want to focus particularly on functions 4 and 5. When used for these functions, Wimsatt says that the false model is used as a "template" that is used to either factor out larger effects in order to capture effects that are too small to be seen, or for estimating parameters that are not themselves included in the model.

The discussion in Wimsatt (2007) involves a detailed study of the linear linkage model developed by Thomas Hunt Morgan in the early twentieth century. Wimsatt gives several examples of cases where deviations from the predictions of the linear linkage model were used to postulate causal factors that were not being taken into account in the model. This use of the model would fall under function 4 (Wimsatt 2007, 106-111). He also discusses a case where deviations from the

predictions of another model, the Haldane mapping function, is used to estimate the value of a parameter that is not contained in the model itself (Wimsatt 2007, 120).

I want to emphasize again that functions 4 and 5 for models as described by Wimsatt is active, not passive, de-idealization. The false model is used *directly* to produce evidence that can then be used to extract information about the system or phenomenon of interest. It is not being used merely as a heuristic—rather, the model itself is used to produce the observations. Wimsatt provides a very good example of these uses of false models, but one might get the impression that the way in which models are used here is relatively rare in science. This impression, however, is mistaken—there are at least some sciences where this is the primary way in which progress is made. Most of our knowledge of the interior of the earth, for example, is the result of the application of this method.

Perhaps one of the reasons that this method has not gotten the attention it deserves is that it raises some rather difficult issues with regard to justification. There is, first of all, a circularity worry. Suppose I create an initial model, and then I study the deviations from this model. These deviations are then taken to be evidence for, say, causal factors that must be taken account in the model. I then add these further causal factors, and improve the model. Perhaps I then investigate further deviations from the predictions of this new model, and try to make further improvements to the model. If, however, the wrong initial model was used, then the deviations might not reflect any real causal factors after all—they might turn out to have been illusory, in which case the research program would have been going down a "garden path".[4] So one thing you would want to be careful about is that if a false model is being used to create new observations, the model ought to be *false in the right way*—the deviations

[4] This is George Smith's term. See Smith 2002.

from the model ought to be ones which actually will tell us true things about the system, or at least point us in the right direction. You would then expect that there might be some norms for models if they are being used in this way. I will discuss such possible norms below. I will now turn to a discussion of earth reference models—models that I believe are used in the way I have described.

3. Earth Reference Models

The earth reference models that I have mentioned in this paper are idealized models of the mechanical properties of the earth's interior. If the interior of the earth is taken to be elastically isotropic, then the mechanical properties of each point in the earth's interior can be characterized by three variables: density, and two parameters that express the elastic properties of the medium, usually incompressibility and rigidity. If the earth is taken to be spherically symmetric, that is, mechanical properties of the earth are taken to depend only on the distance from the center of the earth, then the mechanical properties of the entire earth can be represented completely in terms of three functions of radius. For such an idealized earth, expected travel times for various types of seismic waves can be calculated. In the 1930's, spherically symmetric models of the mechanical properties of the earth's interior were determined by constructing idealized earth models and comparing expected travel times for such models with actual travel times of seismic waves. There was a remarkable agreement between the earth models produced by the two main groups working on earth models at the time, one involving Harold Jeffreys and Keith Bullen, and the other involving Beno Gutenberg and Charles Richter (Bullen 1975). The methods used here were hypothetico-deductive—that is, the models were postulated

as hypotheses about the earth's internal structure, and they were compared directly against observations of travel times.

In the 1960's, the fortuitous confluence of digital computing technology with a couple of the largest earthquakes ever recorded made possible the recording of normal modes of oscillation of the earth. Normal mode frequencies can be calculated for an idealized, spherically symmetric earth, and models that incorporate normal mode frequency observations were built starting in the 1960's. Further, advances in computing allowed geophysicists to develop Monte Carlo methods in which earth models were generated randomly by computer and tested against observations (e.g., Press 1968). Some of the models that agreed with observation were significantly different from the other models that had been postulated at the time, and these studies led to worries about the possibility of radically different models being consistent with observations. Work by the geophysicists George Backus and Freeman Gilbert (Backus and Gilbert 1967, 1968, 1970), which tried to address this non-uniqueness problem, showed that limits could be put on the degree of non-uniqueness of earth models, but only under the assumption that the functions relating earth structure to observations of normal mode frequencies were linear, an assumption that was known to be false.

According to the geophysicist Keith Bullen (1974), a committee was set up in 1971 for the construction of a "Standard Earth Model". The reason given for the construction of this new model is that a large amount of new data had been collected since the Jeffreys-Bullen and Gutenberg-Richter models had been constructed, and individual geophysicists had been incorporating this new data in different ways. This had led to a "great untidiness in the presentation of numerical seismological results". In the mid-1970's, several teams of geophysicists began to develop earth models with

the goal of coming up with a standard reference model. In 1981, this process culminated with the development of the Preliminary Reference Earth Model, which is still being used to this day, although there are now several other alternative models that are used as reference models as well.

Earth reference models, such as PREM, are used in many ways, but what is most distinct about their use from an epistemological point of view is that they are best thought of in terms of functions 4 and 5 in the taxonomy of functions of idealized models described by Wimsatt. They are used, that is, for detecting phenomena that would otherwise be masked or be too small to be seen, or for estimating magnitudes of parameters that are not included in the model.

These two uses can be seen quite clearly in the way in which PREM has been used for the construction of three-dimensional models of the interior of the earth, that is, models that are no longer simply spherically symmetric, but express the mechanical properties of the earth's interior in terms of three spatial variables. Most of these models are based on observations of travel times of seismic waves. They are not, however, constructed by simply constructing a model and comparing it with actual travel times of seismic waves. The observations used are usually travel time residuals—that is, the deviations from the travel times predicted by a reference model such as PREM. The three-dimensional model constructed is then a linear perturbation of a one-dimensional reference model, such as PREM (Ritzwoller and Lavely 1995). Thus, the deviations between the observations predicted by PREM and actual observations are being used to identify three-dimensional features of the earth which are not in PREM itself, and to measure parameters that represent mechanical properties of such additional features.

4. Possible Norms for Reference Models

I now want to think about possible norms that might govern the use of reference models, keeping in mind Wimsatt's functions 4 and 5: detecting phenomena that would otherwise be masked or be too small to be seen, or for estimating magnitudes of parameters that are not included in the model. Reference models are being used to produce new observations through an analysis of the deviations of actual observations and predicted observations of the model. These observations are then used to eventually arrive at a better picture of the earth's interior. In order to be useful in this process, the models are idealized—that is, they are false, but they must be false in the right way. What is "false in the right way", though? There are, I think, two primary norms. First, they must be simple in such a way that they can be utilized easily in this process of producing further observations. Second, they must somehow reflect the physical situation, in such a way that deviations between what they predict and actual observations actually have some kind of physical significance.

Here is an example of how the first norm played into the development of PREM. In the mid-1970's, there were several teams of geophysicists working on different earth models towards the development of the standard reference model. One such model was a "parametrically simple earth model" (Dziewonski, Hales and Lapwood 1975). This spherically symmetric model represented the mechanical properties of the interior of the earth in terms of a piecewise continuous function, where most of the pieces were low-order polynomials. There is, of course, no reason to think that the mechanical properties of the earth are truly distributed in accordance with low-order polynomial functions. There are, however, advantages to this kind of representation. For example, the "travel times of body waves and their derivatives would always vary smoothly as a function of distance on a particular branch of a

travel time curve." (Dziewonski, Lapwood, and Hales 1975, 12) As I mentioned above, one-dimensional reference models are often used for the construction of three-dimensional earth models using travel time residuals as observations. A model in which the predicted travel times varied smoothly as a function of distance would be easier to compute residuals for. This would not only be useful for the construction of three-dimensional models, but also for other investigations that require the use of travel time residuals, such as the location of seismic sources. Ultimately, the representation of large sections of the interior of the earth in terms of low-order polynomials was adopted into PREM (Dziewonski and Anderson 1981) as well.

The other norm is, I think, more complicated. Reference models must be false, but they must be false in a physically meaningful way. Often, what this means is that reference models will not be the best fit model empirically. One of the geophysicists involved in the construction of PREM, Adam Dziewonski, discusses this consideration in a later paper which considers the possibility of constructing a new reference model:

> A reference model, in a modern sense, is one which satisfies more than just one class of seismological or geophysical observations—like, for instance, travel times of body waves. It should constitute a common basis of reference for all the different studies concerning the earth. […] This strategy seeks a model which has to be physically meaningful—as opposed to an empirical one, which could achieve good results at reproducing a narrow range of observations rather than explaining them. (Morelli and Dziewonski 1993, 179)

What is meant here by "physically meaningful" is that deviations from what the model predicts and what actual observations show give us useful information about the interior of the earth.

Exactly what "physically meaningful" means could depend on the specific ways in which the reference model is being used. For example, if the reference model

is used in the construction of three-dimensional models of the earth, it would ideally correspond to the lowest order term in a spherical harmonic expansion of the normal modes of the earth. Deviations from such a model would contain information about higher-order modes that would be indicative of finer three-dimensional structure. However, if a reference model is going to be used for many different purposes, a more general notion of "physically meaningful" might have to be used. This is a complicated matter, on which further work needs to be done. Here, however, I would like to point out the connections between the use of reference models and some recent work on scientific methodology.

5. Turning Data into Evidence

In this final section, I want to discuss connections between the way in which reference models are used in geophysics with some recent work on scientific inference by George Smith (2002)[5] and William Harper (2011). Both Smith and Harper have done extensive work on Newton, and they both emphasize the role of Newton's Fourth Rule for Philosophizing in Newton's methodology:

> In experimental philosophy, propositions gathered from phenomena by induction should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions.
> (Newton 1999, 796)

The Fourth Rule of Reasoning says two things: that we should rule out hypotheses in favor of propositions that are gathered from phenomena, and that we should provisionally take such propositions to be either exactly or very nearly true. Smith

[5] George Smith has, himself, written on earth models (Smith 2007), including PREM, although his focus is on the period in geophysics before the construction of PREM, and not on the uses of PREM and other reference models.

(2002) argues that Newton's methodology involves taking such propositions gathered from phenomena to be provisionally true so that deviations between what you would expect the phenomena to be like, given that the propositions are true, and what the phenomena are actually observed to be like, can be found. These deviations are then taken to be new phenomena that require explanation. Both Smith and Harper refer to this process as "turning data into evidence". They both reject a simple hypothetico-deductive picture where there is a hypothesis, and this hypothesis is supported (or rejected) by data. Instead, certain propositions are needed in order to extract phenomena from raw data—to "turn data into evidence".[6]

The parallel with the use of reference models is obvious. Reference models play the role of propositions gathered from phenomena. Expected observations for these reference models are calculated as if the reference models were true, and then deviations between these expected observations and actual observations are either taken to be indicative of further causal factors, or these deviations are used to try to measure further parameters that are not captured in these models. Reference models are being used, in other words, to turn data into evidence. "Turning data into evidence" is another term for what I have been calling active de-idealization.

If this is, indeed, an accurate picture of a significant way in which science is done, then it might be useful to think about the norms that govern this methodology. If one of the aims of building models—or, more generally, theorizing—is to enable active de-idealization, then we might expect the norms that are required here to be different from those that would govern a more standard picture where models or theories are constructed without active de-idealization in mind. For example, there might be a norm for simplicity that is driven less by notions about the connection

[6] See Miyake 2013 for a more detailed discussion.

between simplicity and truth, or by simple tractability considerations, and more by the fact that models that are simpler in certain ways can more easily be put to use in producing further observations. There might also be a fairly complicated norm for "physical meaningfulness"—one that requires a model to yield deviations that would tell us something about a system or phenomenon of interest.

Now, one notable difference between Newton and earth modelers is that earth modelers already have fairly good ideas about what "physical meaningfulness" amounts to when building earth models, although they might not, by any means, have a complete picture. On the other hand, the whole difficulty for Newton was coming up with a background theory that would allow him to differentiate between what is "physically meaningful" and what is not. Thus, one might think that what I have to say here about reference models does not easily apply to the case of Newton. On the contrary, however, I believe a detailed examination of the use of reference models could, itself, be a useful reference against which to compare the difficulties faced by Newton and others in various important episodes in the history of science.

REFERENCES

Backus, George and Freeman Gilbert. 1967. "Numerical Applications of a Formalism for Geophysical Inverse Problems." *Geophysical Journal of the Royal Astronomical Society*,13:247-276.

Backus, George and Freeman Gilbert. 1968. "The Resolving Power of Gross Earth Data." *Geophysical Journal of the Royal Astronomical Society* 16:169-205.

Backus, George and Freeman Gilbert. 1970. "Uniqueness in the Inversion of Inaccurate Gross Earth Data." *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, Vol. 266, No. 1173:123-192.

Bullen, Keith. 1974. "Introductory Remarks on Standard Earth Model." *Physics of the Earth and Planetary Interiors*, 9:1-3.

Bullen, Keith. 1975. *The Earth's Density*. Chapman and Hall.

Dziewonski, Adam, and Don Anderson. 1981. "Preliminary Reference Earth Model." *Physics of the Earth and Planetary Interiors*, 25:297-356.

Dziewonski, Adam, Anton Hales, and E. R. Lapwood. 1975. "Parametrically Simple Earth Models Consistent with Geophysical Data." *Physics of the Earth and Planetary Interiors*, 10:12-48.

Harper, William. 2011. *Isaac Newton's Scientific Method: Turning Data into Evidence about Gravity and Cosmology*. Oxford University Press.

McMullin, Ernan. 1985. "Galilean Idealization". *Studies in History and Philosophy of Science*, Vol. 16, No. 3, 247-273.

Miyake, Teru. 2013. "Essay Review: Isaac Newton's Scientific Method". Philosophy of Science 80, No. 2, 310-316.

Morelli, Andrea and Adam Dziewonski. 1993. "Body Wave Traveltimes and a Spherically Symmetric P- and S-Wave Velocity Model". *Geophysical Journal International*, 112:178-194.

Morgan, Mary, and Margaret Morrison. 1999. *Models as Mediators*. Cambridge University Press.

Newton, I. (1999). *Mathematical Principles of Natural Philosophy.* (I. B. Cohen and A. Whitman, Trans.). Berkeley: University of California Press.

Press, Frank. 1968. "Density Distribution in Earth." *Science*, Vol. 160, No. 3833:1218-1221

Ritzwoller, Michael, and Eugene Lavely. 1995. "Three-Dimensional Seismic Models of the Earth's Mantle." *Reviews of Geophysics* 33, 1:1-66.

Smith, G. E. 2002. "The methodology of the *Principia*". In I. B. Cohen and G. E. Smith, (Eds.), *The Cambridge Companion to Newton*. Cambridge University Press.

Smith, G. E. 2007. "Gaining Access: Using Seismology to Probe the Earth's Insides". Available for download at: http://www.stanford.edu/dept/cisst/events0506.html

Weisberg, Michael. 2013. *Simulation and Similarity*. Oxford University Press.

Wimsatt, William. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Computation and Scientific Discovery? A Bio-inspired Approach

Ioan Muntean[1]

[1] The *Reilly Center for Science, Technology and Values*, University of Notre Dame, Notre Dame, IN 46556
imuntean@nd.edu

Abstract

Philosophers argue that scientific discovery is far from being a rule-following procedure with a general logic: More likely it incorporates creativity and autonomy of the scientist, and probably luck. Others think that discovery can be automatized by some computational process. Based on a concrete example of Schmidt and Lipson Schmidt and Lipson (2009), I argue that the bottom-up discovery is computable and that both aspects of creativity and autonomy can be incorporated. The bio-inspired evolutionary computation (genetic algorithms) are the most promising tool in this respect. The paper tackles the epistemology of applying a evolutionary computational and genetic algorithms, to the process of discovering laws of nature, invariants or symmetries from collections of data. Here i focus on more general aspects of the epistemology of evolutionary computation when applied to knowledge discovery. These two topics: computational techniques applied in science and scientific discovery taken separately are both controversial enough to raise suspicions in philosophy of science. The majority of philosophers of science would look with a jaundiced eye to both and ask whether there is anything new to say about discovery and computers in science. This paper is a first stab to the philosophical richness of computational techniques applied to the context of discovery. I discuss the prospect of using this type of computation to discover laws of nature, invariants or symmetries and appraise their role in future scientific discoveries.

Is scientific discovery an algorithmic process?

I argue in this paper for a deeper connection between bio-inspired computation and the process of scientific discovery. Based on new concrete results of Schmidt and Lipson 2009, I infer here some epistemological consequences for using evolutionary computation in scientific discovery.

Knowledge is central to virtually all advanced forms of life; discovery and learning characterize us as a species as well as other higher order animals. We discover in order to survive and adapt. Science is just another specific form of knowledge in which data and experiments play a fundamental role in conjecturing hypotheses about the world. If discovery is probably intrinsically linked to our evolution as a whole, scientific discovery played a central role only in the evolution of humanity in the last four centuries or so (a good turning point is the work of Francis Bacon and its influence during the "Scientific Revolution").

How do we infer laws and generalizations from data? How do we discover new models and theories? Are creativity and autonomy of scientists major cognitive faculties that define and shape science, or, on the contrary, is scientific discovery just a process of following rules, methods and algorithms? The nature of scientific discovery, together with, arguably, artistic creativity, moral decision making and religious experience are among those faculties that define us as humans better than anything else.

These fundamental questions about the nature of scientific discovery are germane to the discussion of artificial scientific discovery. As I link the process of discovery to human life as a species, it is germane to investigate philosophically the paths to an artificial process of scientific discovery. Can we create machines that would perform activities deemed by many as "human-only"?

The broader scope of this paper is to investigate the possibility of a cooperation between the human scientist and the artificial discoverer. I based my argument on a specific

Two approaches to scientific discovery

For the purpose of this paper, the scientific endeavor can be divided between the context of discovery and the context of justification. The distinction can be traced to H. Reichenbach's early works but it is very clearly expressed in Reichenbach (1949). After introducing the infamous distinction, Reichenbach discussed the reliability of a logic and epistemology of discovery. Epistemology is a rational reconstruction of a thought process. In a common interpretation, there is no epistemology of discovery, which is basically a subjective and irrational process: P. Duhem, E. Mach, K. Popper, R. Carnap, C. Hempel, or R. Brainwaite for different reasons deemed discovery as irrelevant when compared to the context of justification. The iconoclastic view of scientific discovery as a "happy guess" or "mystic presentiment" is discussed in Koestler (1959). In a different key, M. Curd and Th. Nickles interpreted Reichenbach's discovery-justification distinction as not excluding an epistemology of discovery. There is an epistemology of discovery, with or without a logic of discovery. So epistemology is much a broader area than logic in this specific framework.

For both these contexts it is relevant to ask this question: is science based on deductive logic, induction or on heuristics? A similar question can be asked about the nature of discovery: is scientific discovery algorithmic, nearly algorithmic or, on

the contrary, is it non-discursive, not re-constructible, non-reproducible, singular, a "Eureka"-like mental episode? Is discovery merely a psychological process with no epistemological significance (when compared to the process of justification, for example)?

There are perhaps two main programs in the philosophy of scientific discovery. First, there is a strong program aiming to formulate a general logic for scientific discovery, to encompass all scientific discoveries under one formalism Simon (1973); Hanson (1958). The connection proposed by Langley, Simon, Bradshaw and Zytkow (Langley et al., 1987) between discovery and the heuristic search procedure falls under this strong program. But this strong program felt in disgrace for several reasons and was replaced with a weaker program that gives up the idea of a formal and general *logic* of scientific discovery and tackles the *epistemological* aspects of particular discoveries Nickles (1980b,a); Meheus and Nickles (2009).[1] Here epistemology can be both descriptive and normative and more attention is paid to non-formal and non-logical epistemological aspects of discovery: heuristics, search, risky generalizations, etc. This weak program is more sensitive to the specific conditions of the discovery and of the specific nature of the discoverer. One can ask two questions:

> **(1)** *How do individual scientists, with their limited cognitive faculty, discover new scientific theories? By following a set of rules or by sheer creativity?*
>
> **(2)** *How new theories can be discovered by scientists aided by computers, by Artificial Intelligence systems, or any system other than individual scientists?*

The descriptive epistemology of scientific discovery can answer **(1)** by a careful analysis carried within history of science. Here the discoverer is an individual—the lone genius of Kant, or any scientist experiencing the "Eureka" moment of discovery. We face here a "dilemma of explanation" if we have a theory about scientific discovery as algorithmic Nickles (1980b); Wartofsky (1980):

> **(3)** DILEMMA OF ALGORITHMIC EXPLANATION*: The dilemma is then: either the theory succeeds, and the concept of discovery is explained away, or reductively eliminated—or the theory fails, and discovery remains unexplained.*

I emphasize here the novelty of question **(2)**. First, it does not have a complete answer in the history of science, because the computer-aided scientific discovery or discoveries made by large teams of scientists have a shorter history–when compared to scientific discoveries made by individuals. When the discoverer is a collaborative team, a whole scientific communities, a team working with computers, or a set of computational processes, or all these working together, rationality and creativity may well have radically opposite meanings. The answer to **(1)** does not entail an answer to **(2)**, and vice-versa. Communities, computers or other entities may discover scientific laws, patterns, or theories by an altogether different mechanism than human scientists do, with or without explaining away creativity.

This paper aims to answer **(2)** and show in what sense there is "a third way" in Wartofsky's dilemma **(3)**. The way in which computers and artificial intelligence are used in science may elucidate the normative part of this epistemological approach, but we do not need to equate computational techniques with rational agents, machines, number crunching devices, etc. I do not identify rationality with logic, irrationality with creativity, or machines with logic and creativity with humans only. When used in the scientific discovery, the computational technique comprehends several elements such as: creativity, rule-following procedures, logic etc. I think there is something interesting for philosophers to study about discovery and about computation, taken separately or when computation is directly applied to scientific discovery.

The skeptic against computers used in areas in which human knowledge reigns may raise important questions: Are current computational techniques versatile enough to reproduce, and eventually enhance, the process of scientific discovery? If so, which type of computation is the most promising? And moreover, is this process going to slowly replace humans with machines, even in the process of discover? I reckon that all these questions are attractive from a philosophy of science point of view. It is even more contentious whether a computational process can discover solutions to problems that humans (alone) cannot discover.

In focusing on the epistemology of scientific discovery and the possibility of its algorithmic reconstruction, the current approach is more local and partial: I focus on a specific bottom-up approach to discovery: inferring invariants and laws of nature from large sets of data, and on a specific type of computation: the evolutionary computation implemented by genetic algorithms.

The philosophy of computation in science follows the debates on the relation between data, phenomena, models and theories. For the purpose of my analysis, two contexts of computational science are relevant, both inspired by recent discussions on applying science/applied science Morrison (2006); Bod (2006); Boon (2006). (a) The computational technique starts from a scientific theory and move towards the data: here computation is the *application* of a theory or a "top-down" approach. Or (b), computation is a *heuristic* tool that starts from data and builds a theory in a "bottom-up" approach. Each of these two approaches may have their own specific computational turns: computational techniques used in one may or may not be as revolutionary as they seem in the other. Differentiating these two contexts may help the philosopher argue for the novelty of the epistemological aspects of (b) when compared to (a).

Evolutionary Computation and the Bottom-up Approach to Theory-building

On different occasions, philosophers and scientists alike pointed out to a major difference among two types of scientific reasoning (Th. Kuhn, L. Laudan, among others). On one hand, one has the rule-based reasoning in which new theories or models are inferred from a set of rules. The system of abstract rules is used to solve problems. The rules in general are content-neutral and in the ideal situation they can be applied to virtually any new set of data. On the other hand, one witnesses case-based reasoning in science. Th. Kuhn and K. Popper asked incessantly: is science applied by following rules? Exemplars are solutions to previous problems that scientist learn during their scientific education and solve future puzzles based on an "acquired similarity"Kuhn (1962).

[1] For reasons why the strong program failed, see Curd (1980); Laudan (1980).

Scientists try to make a new phenomenon fit to one or more previous phenomena.

A relevant step forward is to show that neither science, nor computation can be reduced to a succession of rule-following procedures. If we restrict computers to rule-following, then there is little chance, if any, that computational techniques can be useful in scientific discovery. Some philosophers of science have analyzed computation as heuristics device in the *discovery* of new theories. Here concrete results are less notable than in (a). Computer scientists try to use algorithms to discover laws of nature, invariants or patterns in data at least since the 1970s: the most known are the packages DENTRAL, EURISKO, GLAUBER, STAHL and BACON Simon et al. (1981); Mitchell (1997); Waltz and Buchanan (2009). They are designed for a theory-building procedure, when the scientists have little or no idea about how the theory is supposed to look like Keller (2003); Galison (1996); Langley (1979); Barberousse et al. (2007); Pennock (2000, 2007). There is a similarity between the Case-Based reasoning suggested by Kuhn and similar AI techniques used in problem-solving Bod (2006). A case-based procedure always retrieves cases whose problem is similar to the problem being solved. The procedure discussed is data-oriented as opposed to rule-based processing. Computers mimic frequently the process of learning, which is not completely based on rules. According to Bod, data-oriented procedures in computers are similar to the way scientists explain new phenomena "by maximizing derivational similarity between the new phenomenon and previously derived phenomena" Bod (2006).

Therefore, neither scientists nor computers follow strict rules, but reuse previous results in order to solve new problems. For Bod, previous patterns of derivations are learned and accumulated, not phenomena in themselves. Rules are always present, but they are complemented with corrections, normalizations, exemplars derivations, adjustments, all stored and reused from previous cases. In context (b), in the data-oriented discovery process, something else is needed than rule-following procesures. This takes us a step towards answering **(1)** and solving dilemma **(3)**. As P. Langley *et al.*, P. Thagard (1998) and L. Darden (1998) have argued, bringing in computation into the discussion on scientific discovery should majorly boost philosopher's interest in discovery. But, as my argument goes, the nature of computation plays a central role in dismissing **(3)** as a false dilemma and answering **(2)**. I show that once we move to a new type of computation, **(3)** is based on some false assumptions if we give up the very restrictive concept of algorithm and adopt a general concept of computation.

Based on the concrete case study (Schmidt and Lipson, 2009), I show in what sense creativity and rationality can in fact go hand in hand in the case of genetic algorithms applied to scientific discovery. The answer lies in the artificial life metaphor used by Schmidt and Lipson. Computational results in this context are still rare, but as my argument goes, this case cuts deeper into the computational epistemology. More concretely, in the following two sections I address these questions:

> **(4)** *What are the epistemological consequences of using evolutionary computation in scientific discovery?*
>
> **(5)** *Is evolutionary computation the appropriate type of computation for the process of discovery?*

Evolutionary Computation

Roughly speaking, computer algorithms were born based on three distinct analogies: algorithms as "formal proofs", algorithms as "learning processes" and algorithms as "searching procedures for optimality". The latter inspired the area of evolutionary computation, as the paradigm for optimality is an organism optimally adapted to its environment.

How is "search" related to "life"? In the 1930s, S. Wright (1932) interpreted a biological species as a system that evolves in time by exploring a multi-peaked landscape heuristic of optimal solutions to a "fitness problem" . The operation of optimization of search which is typically performed by an algorithm can mimic a living organism that over a long period of evolution fits the environment. On the other hand the process of adaptation and evolution is not smooth.

Organisms are subjected to *random* mutations, too. Taken the biomimetic strategy on step forward: Is it a good idea to add randomness to algorithms? There are several types of *stochastic* algorithms each of them being more or less *biomimetic* in their nature. Biomimetic strategies are widely used in robotics and artificial intelligence, but they are almost ignored by philosophers.[2] Are they useful when applied to scientific discovery?

After a serendipitous proposal by A. Turing in the early 1950s, Evolutionary Computation (*EC*) was rediscovered and reinvented at least ten times before the 1980s (Fogel, 1998). The milestone is J. Holland's work (1975). Following Turing and von Neumann, Holland was able to see the potential of using the knowledge on natural adaptation process to improving search techniques and applied the principles of natural selection directly to problem-solving algorithms. One fundamental difference, not available in Turing's time, is that selection occurs better at the level of population, not at the level of individuals.

The elements of a genetic algorithm

Genetic algorithms are iterative procedures of *searching* for the optimal solution to a problem P. They are based on the metaphor of biological processes in which organisms: (a) *non-consciously* adapt to the "environment" P and (b) are selected by a *supraindividual* mechanism such as selection.[3] The question is whether we can generate algorithms in the same way organisms are created through evolution.

Genetic algorithms start from a given number of initial individuals randomly distributed in a given space, called the initial population. The genetic algorithm transforms individuals, each with an associated value of fitness, into a new generation by using the principles of survival-of-the-fittest, reproduction of the fittest and sexual recombination and mutation. Similar to Wright's landscape, the genetic algorithm finds "the most suitable" or the "best so far" solution to the problem by breeding individuals over a number of generations.

The procedure can be stopped by a termination condition: when the sought-for level of optimality is reached or when all

[2] On the concept of biomimetics, see Srensen (2004); Muntean and Wright (2007).

[3] I take here algorithms as abstract, mathematical objects, whereas programs as their concrete instantiation on a machine. A sensitive difference is between genetic algorithms, genetic programming and genetic strategies. See Jong (2006).

the solutions converge to one candidate. The fitness function estimates the fitness to breeding of individuals in accordance with the principle of survival and reproduction of the fittest:

- Better individuals are more likely to be selected than inferior individuals.
- Reselection is allowed.
- Selection is stochastic.

The genetic algorithm ends with a *termination condition* that can be the satisfying of a success predicate or completing a maximum number of steps. The success predicate depends on the user's choice and can be deemed as a pragmatic criterion. The winner is designated at the "best-so-far" individual as the result of the run.

Here is an abstract implementation of a genetic algorithm:

```
[1] produce an initial population of
            individuals

    [2] WHILE `termination' not met do
       [3] evaluate the fitness of all
            individuals

       [4] select fitter individuals for
            reproduction
       [5] produce new individuals

       [6] generate a new population by
       inserting some new
       good individuals and
       by discarding some
       'bad' individuals

       [7] mutate some individuals

    [8] ENDWHILE

[9] Call the individual(s) which satisfy
       the 'termination' condition
       the ``best-fit-so-far''
```

Case study: (Schmidt and Lipson, 2009): Distilling laws and invariants

To show that "algorithmic explanation" and "creativity" are *not* mutually exclusive in **(3)**, I use as an example of computation applied directly to science the result reported in *Nature* (Schmidt and Lipson, 2009). M. Schmidt and H. Lipson have showed how symbolic regression based on evolutionary programming can be used in discovering *natural*, *non-trivial* and *meaningful* invariants in physics.[4] Their algorithm searches over the infinite possible ways of modeling data to find the best and most useful expression available given (i) a set of data; (ii) a termination condition and (iii) a set of evolutionary path. It starts with a set of individuals which can be equations, models and scientific heuristic methods of search—not necessary mathematical objects. Each individual is tested against a bank of experimental data. Many individuals do not make sense mathematically or do not meet some consistency criteria, so they are discharged. Some may fit the data better than others. The software saves these individuals for "breeding", cross-combining a 'father' with a 'mother'. It is claimed that over hundreds of thousands of generations, some extremely fit individuals emerge.

[4]The package is EUREQA, a software based on evolutionary algorithms Lab (2009).

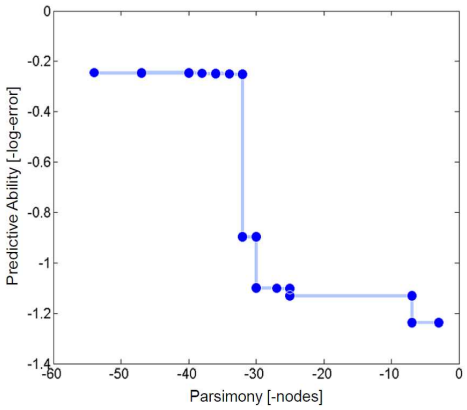


Figure 1: The Pareto front with two "cliffs".(Schmidt and Lipson, 2009, supplementary online materials)

Schmidt and Lipson approached scientific discovery as being data-driven. They started from a set of measured, uninterpreted set of data representing the position, velocity and acceleration of a lab experiment or a virtual system (generated by another algorithm). The method used, the "symbolic regression", is not new at all, but here the program searches for both the form and the parameters of an equation that model a given set of experimental data. They have discovered not only analytic functions from empirical data, but structures which are highly relevant to physical sciences: Hamiltonians, Lagrangians, laws of conservation, symmetries, and other invariants.

Schmidt and Lipson adopted the balance between two objectives: the predictive power and the complexity/parsimony of each candidate. By calculating the "Pareto front" of the dependence predictive ability versus parsimony, Schmidt& Lipson found that there are two cliffs where predictive ability jumps rapidly at some relatively small increase in complexity.

The Epistemology of Discovery with Evolutionary Algorithms: Risks and Advantages

One knee-jerk reaction to applying computation to science is: what is so philosophical about (yet) another tool used by scientists? Although we are nowhere near an "end of computation", the philosopher would not directly infer from its success, its epistemological relevance. Many scientific tools are successful in science, but philosophically inept, and *vice versa*. Although not yet successful, I claim that this case study is worth of a philosophical scrutiny as it sheds some

light on some concepts such as: creativity, rule-following, knowledge production, etc. The procedure addresses some very general epistemological issues of scientific discovery. The knowledge-production in this case study uncovers interesting aspects of the scientific discovery. I frame the following epistemic "aspects" both as problems and as novel features of the scientific discovery based on evolutionary algorithms. The direct application of evolutionary computation to scientific discovery shows how productive bio-inspired algorithms can be. The most attractive feature of evolutionary computation is its ability to "explore" the logical space of solutions, even those which remains unconceived to the mind of the scientist. But the whole process is not totally automatized and the algorithm is not fully autonomous. The human scientist imposes her own meta-rules on the algorithm. On the other hand, because every solution is a model better or worse adapted to data, the bio-mimetic aspect of this example is clear: scientific models adapt to the data and create populations of solutions such that each individual contributes to the adaptation function of the population. After running the algorithm as suggested by Schmidt and Lipson, the scientist is able to explore the "tip of the iceberg", i.e. the best adapted in so far individual from a multitude of previous generations of solutions. The unconceived alternative models, although not direcly present in the final solution did influence it if they were part of the intermediate generations of solutions. I relay the epistemological aspects of the genetic algorithms used in scientific discovery to the various aspects of artificial life. A stronger connection, not endorsed here, would connect knowledge in general to evolution, the are being the evolutionary epistemology. The main part of my argument is that the face of scientific discovery "as we know it" may change radically once evolutionary computation is involved in the process of discovery. I list here several aspects of this "upward epistemology" that is still nascent but very enticing philosophically.

Stochasticity versus scrutability of solutions

Genetic algorithms can be stochastic or not, depending on the mutation operator occurring in step (7) or by selecting the individuals for reproduction in step (5) (in Table 1). An algorithm becomes deterministic if exactly one parent is *identically* reproduced or if two parents are combined without adding or losing information based solely on their fitness. Genetic algorithms are stochastic in two major respects: both the operation of selection and reproduction are random. That means the results (offspring) are not direct results of the input data (the parents).

The crossover operator takes two individuals, the parents, and produces two new individuals, the offspring, by swapping substrings of the parents. Randomly choosing two parents to mate or randomly deleting or adding information from the parents will make the algorithm stochastic. Mutation is a background redistribution of strings to prevent premature convergence to *local* optima.

Weak individuals may survive "by luck" and fit individuals may not be drawn to reproduce. The advantage of a random mutation is that at least some populations, ideally a few only, could escape the traps which deterministic methods may be captured by, and end up with an unexpected and novel result. For very complex problems, this biomimetic procedure can output results which are definitely not accessible to deterministic algorithms if a delicate balance between the mechanism of selection that decrease variation and those that increase variation (mutation) has been achieved.

Because the scientist can control this mutation operator and its frequency, the output of such a discovery algorithm is not traceable by humans. At the limit, the solution of such an algorithm may be inscrutable to humans. It is also the case that for any run, because of the stochastic element, the best individuals are not guaranteed to be selected, and the worst are not eliminated. One can say that the algorithm favors the best and marginalizes the unfit. The selection is not entirely "greedy" in the search space. We do not need to associate creativity to such a random process. As I show before, human element is not totally eliminated in this case. The creativity is blind in this case, similar to mutation in biological populations.

Rules, laws and metarules

The evolutionary algorithms do not follow a set of rules in respect of the discovery of new laws or invariants. As the case study suggests, the process of discovery is here ruled by the metarules of evolution as well as the method used to decide about the fitness function and the termination condition.

For simple laws and invariants, genetic algorithms are easily outrun but Turing machines. But given the complexity of current science, deterministic algorithms may well be worn out as aiding tools to optimality. Although this may sound speculative, let us assume that science evolved toward increasingly complex representations. Maybe the good-old-days of simple, beautiful laws of nature are gone. What if were not going to encounter beautiful laws such as:

$$F = ma; F = k\frac{m_1 m_2}{r^2}; E = mc^2; R_{ij} - \frac{1}{2}g_{ij}R = 8\pi G T_{ij}$$

anywhere down the road? For the time being, weve been lucky enough that our best laws of nature could have been fit on a "T-shirt", as it were. How do we discover more and more complex laws of nature? We are limited by conceivability and our limited resources to recognize patterns and regularities may become overtaken by the increasingly complex set of data. Time in which we could deduct laws from phenomena without any epistemic extenders may be over. More and more complex data are collected. Cosmologists, neuroscientists, sociologists, political scientists do not have the luxury to infer their laws from laws as simple as Newton's or Einstein's. What if, from now on, the would-be laws of nature wont fit even a football banner? We need to brace up for more and more complex scientific representations...

Social science, biology, suggest that we may want to drop completely the ideal of laws of nature in their simplest and purest form. In some historical cases, pre-existing theories and the accompanying mathematics were not "already there" when a major discovery in science occurred: contrast this with the received view on the "unreasonable effectiveness of mathematics". We may even need to reconsider the concept of universal laws of nature, existing independent of the way we collect and simulate data.

Now, here is a brighter perspective. Even if the good old days are bygone, there are new ways of coping with increasing complexity in the form of invariants, regularities, laws of nature and alike. Distributive knowledge in science is a tempting idea. Science made by communities of scientists, labs, research programs may steadily replace science

made by individuals. The other possible path suggested by Humphreys is a collaborative work between computers and humans. Maybe we have to face the fact that science is getting closer to the limits of our knowledge, we as limited individual brains. Philosophically put, science is getting closer to the conceivability limit of possibilities.

Triviality versus meaningless

In Schmidt and Lipson's approach, there is a problem of triviality and meaninglessness of solutions. For almost any set of empirical data there are uncountable invariants or conserved quantities, some of them being trivial, some being meaningless. The main task in this case is to find a non-trivial invariant of the system that also can be interpreted as having a meaning. Schmidt and Lipson proposed a criterion based on decomposability: the candidate equations should predict connections between dynamics of the *subcomponents* of the system. This is done by pairing the variables and looking for natural behaviors of parts of the system. More precisely, the conservation equations should be able to predict connections among derivatives of groups of variables over time, relations that we can also readily calculate from new experimental data. Ultimately, their procedure was able to infer the *optimal* form of the double pendulum Hamiltonian by avoiding trivial and meaningless solutions. Schmidt and Lipson included a human decision maker in their algorithm who stops the search process at certain time and imposes the constraints of the symbolic regression such as: "naturalness", "interestingness" or "meaningfulness".

Interpretation versus understanding

Bootstrapping can also be used to infer laws for more complex systems. Results about simpler systems can be used to infer equations for more complex systems. From a statistical analysis, Schmidt and Lipson inferred that terms that are frequently used and are more complex have also *meaning*. For example, trigonometric terms represent potential energy, squared velocities are associated to kinetic energy. The main claim of Schmidt and Lipson is that these terms are ready for a human interpretation:

> These terms may make up an 'emergent alphabet' for describing a range of systems, which could accelerate their modeling and simplify their conceptual understanding. [...] The concise analytical expressions that we found are amenable to human interpretation and help to reveal the physics underlying the observed phenomenon. Many applications exist for this approach, in fields ranging from systems biology to cosmology, where theoretical gaps exist despite abundance in data.
>
> Might this process diminish the role of future scientists? Quite the contrary: Scientists may use processes such as this to help focus on interesting phenomena more rapidly and to interpret their meaning Schmidt and Lipson (2009).

The outcome of such an algorithm can help *in the future* with understanding scientific results which are not strictly speaking discovered by humans. The operation of distilling laws from data does more than generating symbols, be them complex expressions of conserved quantities or equations. Similar to numerical simulations, "the results are not automatically reliable" and more effort and human expertise is needed to decide what results are reliable and which are not (Winsberg, 2009). But in this case the computation is more than a tool or a technique because it makes the results intelligible to the human scientist and the question whether the method can be truly creative is up for grabs.

Path dependency versus global solutions

Genetic algorithms compensate some of their drawbacks by their effectiveness in global search. Remember that they maintain a population of solutions which are constantly updated with fitter new individuals and hence avoid local optima. For a certain complexity of the search space, a genetic algorithm has a better chance to find the global optimum. This changes radically the epistemological aspects of genetic algorithms. They are very efficient in solving "hard problems" where little or nothing is known about the sought-for structure and when discovering new structures trumps the process of evaluating existing knowledge.

The case study underscores well this problem of any evolutionary computation: its path dependence. Even the non-trivial and meaningful solutions are not unique! The procedure does not produce a single set of solutions, but a set of *candidates* for the analytical solutions. It is known that any complex problem has a number of local maxima in the landscape of solutions with different fitness values. At different runs of the simulations, different populations can converge to different maxima. The human discoverer will always reach only one solution, whereas a set of genetic algorithms running on the same initial population will end up with different optimal solutions. This is a direct consequence of the fact that similar to biological evolution, the process is non-deterministic. As it was recently argued, this leads to a non-modular functionality of the algorithms and hence to a limited understanding of the operations (Kuorikoski and Pyhnen, 2013). The only aspect which is etymologically accessible to the scientist is comparing results and deciding the best fit. But the way we achieved that results is inscrutable to the scientist. Previous generations and the evolution itself is in many cases too complicated to follow or alternatively, too stochastic to constitute a justification *per se*. As we cannot trace the proof of the algorithm and replicate it, this is in direct analogy with the way we can run the tape of life and every time a different rational agent will emerge as the "better-to-fit". The principles of recombination, selection, and mutation are basically "operators" in the algorithm to generate new individuals.

Turing versus non-Turing; abstraction versus implementation

This aspect is more speculative and reflects a general attitude towards computation in general. Why is evolutionary computation so special? Some theoretical results suggest that evolutionary Turing machines may are more expressive than Turing machines—at an abstract level.[5] The so-called "Turing Evolutionary machine" is more expressible than an ordinary Turing machine, and its output can converge to the output of an universal Turing Machine. More importantly, the evolutionary Turing machine can solve the TM-unsolvable halting problem using non-algorithmic means (Eberbach, 2005). Generalizing computation to a non-Turing aradigm would

[5] I will follow here mainly Eberbach (2005). See also Pudlk (2001).

provide novel and unexpected epistemological results. Unlike Turing machines, the theory of Evolutionary Turing Machines is relatively unknown to the philosophical community. Eberbach has showed that evolutionary computation can be non-algorithmic, can evolve non-recursive functions and that an evolutionary Turing machine can solve the TM unsolvable halting problem of a UTM. "They are specific metaalgorithms (i.e., algorithms operating on other algorithms) with no restriction on their domain and some (rather historical) restriction on evolutionary algorithms that they have to be probabilistic, population-based, and using fitness function." Eberbach (2005). Practical implementations of evolutionary computation are approximations of Turing machines and they are heavily restricted to time and resources of concrete implementations.

Conclusion

With its "upward epistemology", evolutionary computation applied to discovery is a promising new tool for future scientific projects. Evolutionary computation and genetic algorithms in particular, anticipate the way scientific methodology and knowledge may look in a couple of decades. And the philosopher of science cannot wait for the foreseeable moment of the informational singularity when artificial intelligence will compete with humans. My humble philosophical prediction is that evolutionary computation, or some more "evolved" offspring of it, will be there at the "singularity" party - if there shall be any.

References

Barberousse, A., Franceschelli, S., and Imbert, C. (2007). Cellular automata, modeling, and computation.

Bod, R. (2006). Towards a general model of applying science. *International Studies in the Philosophy of Science*, 20(1):5–25.

Boon, M. (2006). How science is applied in technology. *International Studies in the Philosophy of Science*, 20(1):27–47.

Curd, M. (1980). The logic of discovery: an analysis of three approaches. In Nickles, T., editor, *Scientific discovery, logic, and rationality*, volume 56. Springer.

Darden, L. (1998). Anomaly-Driven theory redesign: Computational philosophy of science experiment. In Bynum, T. and Moor, J., editors, *The Digital Phoenix*. Blackwell, Cambridge.

Eberbach, E. (2005). Toward a theory of evolutionary computation. *BioSystems*, 82(1):1–19.

Fogel, D. B., editor (1998). *Evolutionary Computation: The Fossil Record*. Wiley-IEEE Press, 1 edition.

Galison, P. L. (1996). Computer simulations and the trading zone. In Galison, P. and Stump, D. J., editors, *The Disunity of science : boundaries, contexts, and power*. Stanford University Press, Stanford Calif.

Hanson, N. R. (1958). *Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science*. Cambridge University Press.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, second edition bradford books, 1992 edition.

Jong, K. A. d. D. (2006). *Evolutionary Computation*. MIT Press: A Bradford Book, Cambridge MA, 1st edition.

Keller, E. (2003). Models, simulation, and 'Computer experiments'. In Radder, H., editor, *The Philosophy of Scientific Experimentation*, pages 198–215. University of Pittsburgh Press.

Koestler, A. (1959). *The Sleepwalkers: A History of Man's Changing Vision of the Universe*. MacMillan, Los Angeles.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University Of Chicago Press, 3rd (1996) edition.

Kuorikoski, J. and Pyhnen, S. (2013). Understanding nonmodular functionality: Lessons from genetic algorithms. *Philosophy of Science*, 80(5):637–649.

Lab, C. M. (2009). Eureqa.

Langley, P. (1979). Rediscovering physics with BACON.3. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*.

Langley, P., Bradshaw, G. L., Simon, H. A., and Zytkow, J. (1987). *Scientific discovery : computational explorations of the creative processes*. MIT Press, Cambridge, Mass.

Laudan, L. (1980). Why was the logic of discovery abandoned? In Nickles, T., editor, *Scientific discovery, logic, and rationality*, volume 56. Springer.

Meheus, J. and Nickles, T., editors (2009). *Models of Discovery and Creativity*. Springer, 1st edition. edition.

Mitchell, S. D. (1997). Pragmatic laws. *Philosophy of Science*, 64(4 Supplement):S468–S479. Journal Article.

Morrison, M. (2006). Applying science and applied science: Whats the difference? *International Studies in the Philosophy of Science*, 20(1):81–91.

Muntean, I. and Wright, C. D. (2007). Autonomous agency, AI, and allostasis a biomimetic perspective. *Pragmatics & Cognition*, 15(3):485–513. Journal Article.

Nickles, T., editor (1980a). *Scientific discovery, case studies*, volume 60. D Reidel Pub Co.

Nickles, T., editor (1980b). *Scientific discovery, logic, and rationality*, volume 56. Springer.

Pennock, R. T. (2000). Can darwinian mechanisms make novel discoveries?: Learning from discoveries made by evolving neural networks. *Foundations of Science*, 5(2):225–238.

Pennock, R. T. (2007). Models, simulations, instantiations, and evidence: the case of digital evolution. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):29–42.

Pudlk, P. (2001). Complexity theory and genetics: The computational power of crossing over. *Information and Computation*, 171(2):201–223.

Reichenbach, H. (1949). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Literary Licensing, LLC.

Schmidt, M. and Lipson, H. (2009). Distilling Free-Form natural laws from experimental data. *Science*, 324(5923):81–85.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science*, 40(4):471–480. ArticleType: research-article / Full publication date: Dec., 1973 / Copyright 1973 Philosophy of Science Association.

Simon, H. A., Langley, P. W., and Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese*, 47(1):1–27.

Srensen, M. H. (2004). The genealogy of biomimetics: Half a centurys quest for dynamic IT. In Ijspeert, A. J., Murata, M., and Wakamiya, N., editors, *Biologically inspired approaches to advanced information technology*, volume 3141 of *Lecture notes in computer science*, page 496. Springer, Berlin; New York. Book, Section.

Thagard, P. (1998). Computation and the philosophy of science. In Bynum, T. and Moor, J., editors, *The Digital Phoenix*. Blackwell, Cambridge.

Waltz, D. and Buchanan, B. G. (2009). Automating science. *Science*, 324(5923):43 –44.

Wartofsky, M. (1980). Scientific judgement: Creativity and discovery in scientific thought. In Nickles, T., editor, *Scientific discovery, case studies*, volume 60. D Reidel Pub Co.

Winsberg, E. (2009). Computer simulation and the philosophy of science. *Philosophy Compass*, 4(5):835–845.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proc of the 6th International Congress of Genetics*, volume 1, page 356366.

Structural Chaos

Conor Mayo-Wilson

June 12, 2014

Abstract

Philosophers often distinguish between parameter error and model error. Frigg et al. [2014] argue that the distinction is important because although there are methods for making predictions given parameter error and chaos, there are no methods for dealing with model error and "structural chaos." However, Frigg et al. [2014] neither define "structural chaos" nor explain the relationship between it and chaos (simpliciter). I propose a definition of "structural chaos", and I explain two new theorems that show that if a set of models contains a chaotic function, then the set is structurally chaotic. Finally, I discuss the relationship between my results and structural stability.

Climate scientists need at least two types of information to generate forecasts: (1) data about the earth's current climate and (2) a model that describes how the climate changes over time. Thus, there are at least two causes of inaccuracy in climate predictions. First, predictions might may be inaccurate because current climatic conditions are mismeasured or misestimated. Call this **initial conditions error** (ICE). Alternatively, error may arise from an inaccurate model of how the climate changes over time. Call this **structural model error** (SME).[1]

The same remarks apply to predictions about any dynamical system. If one is interested in predicting the evolution of an ecosystem over time (e.g., how population levels of various organisms change), or the behavior of markets (e.g. how prices of various commodities change), or how an epidemic will spread through a city, etc., one needs to identify both the initial conditions of the system and how the system changes over time. So there are likewise at least two sources of error in all these problems.[2]

[1] This distinction is similar to Parker [2010]'s distinction between parameter and model uncertainty.

[2] For a discussion of other sources of error in modeling, see Bradley [2012].

In a recent paper, Frigg et al. [2014] argue that the distinction between SME and ICE is crucial for both scientific practice and policy-making. They claim that, although there are methods that can generate accurate predictions in the presence of both (i) ICE and (ii) chaos, there are no known methods for doing the same with respect to (i') SME and (ii') an analogous notion of "structural chaos", which they call the "hawk-moth" effect.[3] For this reason, Frigg et al. [2014] argue that structural chaos and SME are neglected, but important topics within philosophy of science.

Although they provide an illustrative example and ample computer simulations to suggest structural chaos might be widespread, Frigg et al. [2014] do not define "structural chaos" or investigate its relationship to chaos (simpliciter).[4] This is important because there are many definitions of "chaos", and so there might be many analogous notions of "structural chaos."[5]

Frigg et al. [2014]'s arguments, therefore, raises at least three important questions for philosophers of science, applied mathematicians, and working scientists. First, for each definition of "chaos", what is the analogous concept of structural chaos? Second, what are the relationships among the various notions of chaos (simpliciter) and the analogous notions of structural chaos? Finally, what are the implications of structural chaos for prediction, control, and explanation?

This paper takes a preliminary step with respect to the first two questions. Section one describes some conditions that are used to define "chaos." I focus on topologically mixing systems, which are an important class of chaotic ones.[6] In section two, I define an analogous notion of "structural mixing" that might be used to characterize structural chaos. I then prove that, when a sufficiently rich collection of models contains a topologically mixing function, then the collection is structurally mixing in my sense.

Section three explores the relationship between my results and other

[3]Similar arguments appear in [Parker, 2011].

[4]Frigg et al. [2014] do formally define what they call "closeness to goodness fit." This definition is analogous the the definition of sensitivity to initial conditions, which is generally considered to be a necessary but insufficient condition for chaos. See section one below. At points, they implicitly suggest that structural *in*stability might be the structural analog to chaos. This suggestion is criticized in the section three.

[5]For discussions of definitions of chaos, see [Batterman, 1993] and [Werndl, 2009].

[6]According to Devaney et al. [1989]'s widely-cited definition, a system is chaotic if it satisfies three conditions: (i) it is sensitive to initial conditions; (ii) it is topologically transitive, and (iii) its periodic points are dense in state space. Topological mixing systems are topologically transitive, and under very general conditions, they are also sensitive to initial conditions. Thus, they satisfy two of the three properties that are widely used to define "chaos."

potential characterizations of structural chaos. In particular, I argue that definitions of "structural instability", which are often informally motivated in ways analogous to definitions of chaos, are not clearly analogous to notions of chaos simpliciter. The final section discusses the philosophical importance of my results and answers to the above three questions.

1 Chaos

Popular writings often describe chaos via an appeal to Lorenz's metaphor of the "butterfly effect". Lorenz famously asked whether the flapping of a butterfly's wings in Brazil could cause a thunderstorm in Texas. In general, a chaotic system is often described as one in which small changes (e.g. a butterfly flapping its wing) in the initial conditions of a system can create large changes in its behavior (e.g., storm patterns).

This informal gloss captures only one aspect of standard definitions of "chaos", however. To give more precise characterizations, it is necessary to introduce some definitions. A discrete-time **dynamical system** is a triple $\langle X, d, \varphi \rangle$ where (i) $\langle X, d \rangle$ is a metric space called the **state space**, and (ii) $\varphi : X \to X$ is a **time-evolution** function.[7] For the remainder of the paper, I use the phrases "model", "dynamical function" and "time-evolution function" interchangeably, though of course I recognize not all models in science are time-evolution functions.

For example, a dynamical system might describe the motion of a particle in space. In this case, X is be three-dimensional space; d represents a function specifying the distance between points in three-dimensional space, and φ is a function describing how a particle moves over time. Or X might be the set of vectors specifying the temperature, pressure, and wind velocities at different places in the atmosphere; d would represent how similar two descriptions of the earth's climate are, and φ would represent how the climate changes over time.

How can one use the definition of a dynamical system to capture the notion of sensitivity to initial conditions? Let Δ be a number representing a large distance between states. What counts as "large" can depend upon the state space and one's interests. Say a dynamical system's behavior is **sensitive to initial conditions** to degree Δ if for every state $x \in X$ and every arbitrarily small distance $\epsilon > 0$, there exists a state y within distance ϵ of x and a natural number N such that $d(\varphi^N(x), \varphi^N(y)) > \Delta$. Here, $\varphi^N(x)$

[7] Note that, for simplicity, I assume that the time evolution function φ is constant over time. Not all discrete dynamical systems have this property.

represents the state of the system after N stages if its initial conditions were x. Informally, a system exhibits sensitivity to initial conditions if no matter the true initial state x, there is an arbitrarily close state y such that, if y had been the initial state, the future would have been radically different.

This mathematical definition is the natural way of capturing the above informal description of chaos above, but there are many time-evolution functions that are sensitive to initial conditions in the above sense and yet are hardly "chaotic" in any sense of the word. Consider, for example, the function $f(x) = 2x$ on the state space consisting of all real numbers. Then f is sensitive to initial conditions because if two numbers x and y differ by even the smallest amount, then the result of multiplying them by two repeatedly will cause them to drift apart. That is, $|f^n(x) - f^n(y)| = 2^n|x - y|$ becomes arbitrarily large as n grows. So f is sensitive to initial conditions, but f does not exhibit "chaotic" behavior in the least.

What other conditions might one add in order to characterize "chaos"? It turns out there is no wide agreement, and that several different definitions of chaos are common.[8] Because my aim is to show how three types of questions might be answered (see above), I will not defend a particular analysis of chaos. Rather, I will simply show how to answer the three questions with respect to the concept of "topologically mixing", which plays an important in characterizing chaos (see footnote 5).

A time-evolution function φ is called **topologically mixing** if for any pair of non-empty open sets U and V, there exists a number $N > 1$ such that

$$\varphi^n(U) \cap V \neq \emptyset.$$

for all $n \geq N$. In order to reduce the amount of technical jargon, I will say φ is **chaotic** if it is topologically mixing.

For the reader unfamiliar with topology, ignore the phrase "open set" for now. Just think of U and V as representing sections of state space. If the system begins in some state in U, then the expression $\varphi^n(U)$ represents all possible future states after n many steps of time. For example, suppose the dynamical system describes the movement of a gas molecule in a room. Further, assume that U represents the upper-left quarter of the room and that V represents the lower-right hand corner. Then $\varphi^n(U)$ represents the

[8]For what it's worth, I agree with Werndl [2009] that the vast majority of systems that are agreed to be chaotic are strongly mixing in the sense of ergodic theory. Moreover, I agree with [Berkovitz et al., 2006] that, because strong mixing is one among several logically related concepts of probabilistic independence in the ergodic hierarchy, it is probably most productive to think of chaos as coming in degrees, where different degrees may have different implications for prediction, explanation, and control.

possible positions of the gas molecule after n many units of time if the gas particle started in the upper-left quarter of the room. The above equation says that there is some time in the future such that, from that point onward, there is always a position in the upper-left corner of the room (U) such that, if the gas particle had started in that position, then it would end up in the lower-right quarter of the room (V). A time-evolution function is chaotic if this holds for any regions of state space, which is to say that (in the example) a gas particle that starts in one section of the room can end up in any other section of the room after a sufficiently large period of time.

If topological mixing is taken to be a characteristic of chaotic systems, would would it mean to say that SMEs can lead to "structural chaos"? This is the topic of the next section.

2 Structural Chaos

A dynamical system is chaotic if, when the time-evolution function is held fixed, similar initial conditions can have any future. Analogously, a set of dynamics should be called "structural chaotic" if, when the initial conditions are held fixed, similar time-evolution functions can produce any future. See figure below. To rigorously define "structural chaos", therefore, one needs a metric to quantify how "close" two time-evolution functions are.

Let X^X represent all time-evolution functions for a system with state space X. Depending upon one's interests, there are different appropriate metric quantifying the distance between models (i.e. time-evolution functions). However, clearly there is some relationship between (1) the distance between two models and (2) the distances between their predicted future states after one unit of time. If two models entail that a system, starting in the same initial position, will be in radically different places in a short amount of time, then the models are substantially different.

One demanding notion of closeness requires that two models are close precisely if their values are *always* close. In other words, the distance between two time-evolution functions is the maximum/supremum distance between the models after one unit of time, where the maximum is taken over all possible starting states. In symbols, define:

$$D(\varphi, \psi) = \sup_{x \in X} d(\varphi(x), \psi(x)).$$

Henceforth, I assume that D quantifies the distance between two time-evolution functions, but the results below hold for a variety of metrics.

"Structural mixing" should capture the idea that similar models can produce different trajectories through the state space given the same initial conditions. To make this idea rigorous, I introduce some notation. For any $\epsilon > 0$, let $B_\epsilon(\varphi)$ denote all models within distance ϵ of φ. Next, for any natural number $n \in \mathbb{N}$ and any point $x \in X$, define a map $f_{x,n} : \mathcal{P}(X^X) \to \mathcal{P}(X)$ as follows:

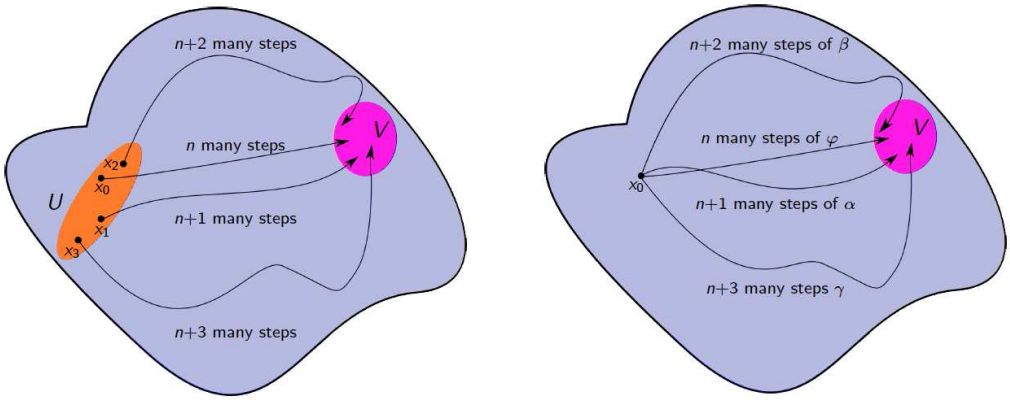
$$f_{x,n}(\Phi) = \{\varphi^n(x) : \varphi \in \Phi\}$$

where $\mathcal{P}(X)$ is the power set of X, i.e., the set of all subsets of X. In other words, $f_{x,n}$ maps a set of time-evolution functions to the set of points they reach after n stages if they are initialized to start at x.

Given a set of time-evolution functions $\Phi \subseteq X^X$ and a particular model $\varphi \in \Phi$, say that Φ is **structurally mixing at** φ if for all $x \in X$, all $\epsilon > 0$ and all non-empty open sets $V \subseteq X$, there is some $N \in \mathbb{N}$ such that

$$f_{x,n}(B_\epsilon(\varphi) \cap \Phi) \cap V \neq \emptyset$$

for all $n \geq N$. In other words, small differences between the estimated model and the true one can lead to divergent predictions *even if one correctly identifies the initial condition.* To reduce jargon, I sometimes say a set of models is **structurally chaotic** at φ if it is structurally mixing.



Topological Mixing Structural Mixing

The concept of structural mixing is the obvious analog of the definition of topological mixing in the previous section. Clearly, different definitions of chaos will generalize to different definitions of structural chaos. Nonetheless, this example suggests a new research program, which consists of three questions. First, for each definition of "chaotic system", what is the analogous

concept of structural chaos? Second, what is the relationship between the various notions of chaos (simpliciter) and the analogous notions of structural chaos? Finally, what are the implications of structural chaos for prediction, control, and explanation?

Given my definition of structural chaos, the second question can be given a precise answer:[9]

Theorem 1 *Suppose φ is continuous and topologically mixing. If X has no isolated points, then X^X is structurally mixing at φ.*

That is, the set of possible time-evolution functions is structurally chaotic if it contains a chaotic model. One might object that this theorem is very weak. According to the theorem, one should worry about structurally chaos if *every* time-evolution function were a plausible description of the dynamics of the system. However, in practice, the set of plausible models is much narrower given existing data, domain knowledge, physical constraints, and so on. For example, if it were $40°C$ in Demascus today, then it would be bizarre if it snowed tomorrow. However, one possible time-evolution function for Demascus' weather will entail that a $40°C$ day will be followed by a snowy day. Thus, one might object that if the class of models is restricted to realistic time-evolution functions, then structural chaos will be rarer.

However, the proof of the above theorem actually shows something much stronger. It shows that, if the true time-evolution function is chaotic and the set of possible time-evolution functions contains all models that are empirically indistinguishable from the true one, then structural chaos will arise. To explain why, I introduce some definitions.

Data sets are always finite. So let $F = \{x_0, \ldots, x_n\}$ by a finite set of states, which represents the observed history of the system so far. Let $\epsilon > 0$ be a small number representing the precision of one's measurement devices. Say two models are **ϵF-indistinguishable** if (1) the values of time-evolution functions are equal for all but finitely many states outside F and (2) the two models are no more than ϵ apart according to D.

Two models are ϵF-indistinguishable if they are, in a very strong sense, indistinguishable given all available empirical data. Why? The first clause entails that the two models are equal on all observed data points, and so there is no way that past data alone can distinguish between them. If two models differ *anywhere*, however, then there are logically possible experiments that can distinguish between them. Namely, if controlled experiments are financially, pragmatically and ethically feasible (which they often

[9]See appendix for a proof.

are not), one can initialize the system to one of the states at which the two models differ and observe the results.

This is where the second clause kicks in. Suppose scientists' measuring instruments and statistical techniques cannot guarantee estimates of the observed states with accuracy better than $\epsilon > 0$. If two models are ϵF-indistiguishable, then second clause guarantees that no information about the current or next state of the system is sufficient to distinguish the models. One might object that small measurement errors are detectable in the long run, especially if the model is chaotic. However, if the true dynamics are continuous and ϵ is sufficiently small, then the second clause entails that no experiment of a feasible length (i.e. time) will distinguish between it and an ϵF-indistinguishable model.

The previous discussion motivates the following definition. Let F denote the finite set of observed states. Say a set Φ of time-evolution functions is **closed under empirical-indistinguishability** if there exists some $\epsilon > 0$ such that if $\varphi \in \Phi$ and ψ is ϵF-indistinguishable from φ, then $\psi \in \Phi$. The above argument is intended to show that, if scientists are strict empiricists, then the set of models that they consider possible ought to be closed under empirical indistinguishability. Theorem 1 is a special case of the following stronger result.

Theorem 2 *Suppose φ is continuous and chaotic. Let Φ be a set of time-evolution functions containing φ. If X has no isolated points and Φ is closed under empirical indistinguishability, then Φ is structurally chaotic at φ.*

3 Structural Stability: Conclusions and Future Research

Readers familiar with chaos theory may find the previous theorem surprising. On one hand, my definition of "structural chaos" seems to formalize the idea that small errors in identifying the model can lead to divergent future behavior. On the other hand, many of the time-evolution functions that lead to "structural chaos" (according to my definition) are *structurally stable* in one or more senses.[10] This is counter-intuitive because structural stability

[10]Suppose $f : A \to A$ and $g : B \to B$ are functions on topological spaces. Then f and g are said to be *topologically conjugate* if there is a homeomorphism $h : A \to B$ such that $g \circ h = h \circ f$. A function $f : A \to A$ is C^r *structurally stable* if there is some $\epsilon > 0$ such that every function within distance ϵ of f in the C^r metric is topologically conjugate to f. C^r structural stability is perhaps the most common definition, but other definitions have a similar logical form, which is discussed in the body of the paper.

is intended to formalize the idea that small changes to the model do not result in large differences in the model's trajectory.

It is best to begin with an example to understand the tension. A paradigmatic chaotic function is the *logistic map* $F_r(x) = rx(1-x)$, where r is greater than about 3.57. It is known that the logistic map is C^2-structurally stable when $r > 4$, and it is structurally stable on an open dense set of values of r between 0 and 4. For this reason, some chaos theorists might claim that small changes to the logistic map will not result in divergent future behavior. However, the logistic map (for $r = 4$) is precisely the example that Frigg et al. [2014] use to demonstrate the impacts of structural chaos. Moreover, if Φ is a set of models that contains the logistic map and is closed under empirical-indistinguishability, then Theorem 2 entails that Φ is structurally chaotic at the logistic map, as the logistic map is topologically mixing. So Frigg et al. [2014]'s and my results seem to be in tension with facts about structural stability.

One possible reason for the tension is that definitions of structural stability almost always assume that the set of models under investigation are well-behaved, in the sense that models are differentiable (perhaps several times) and hence, continuous. In contrast, in order to demonstrate the existence of "structural chaos" in computer simulations, Frigg et al. [2014] simulate discretized functions that are, by necessity, discontinuous. Moreover, if a set of models is closed under empirical indistinguishability in my sense, it will contain discontinuous functions and other "poorly behaved" models. Some may see this as a deficiency in Frigg's and my arguments. Continuity and differentiability are mathematically convenient assumptions, and Ockham's razor or other metaphysical arguments might lead one to accept that the dynamics of real physical systems are continuous. Nonetheless, convenience and simplicity are extra-empirical considerations; a finite sequence of observed states may be consistent with assuming the continuity of the system's time-evolution function, but it does not require doing so. Furthermore, many metaphysical arguments for continuity do not obviously extend to showing that a function is twice differentiable.

However, I will not defend the thesis that physical laws might be discontinuous or non-differentiable. Rather, I discuss the relation between structural chaos (in my sense) and various notions of structural stability in order to illustrate a broader point. Mathematicians, scientists, and philosophers have yet to investigate whether plausible structural analogs of "chaos" are actually in tension with definitions of structural stability. My results show that there may, in fact, be no direct logical inconsistency, and that inconsistency may only arise when additional, substantive assumptions (e.g. conti-

nuity or differentiability) about the dynamics of the system are introduced.

There are two further reasons to question whether standard definitions of "structural instability" are really the appropriate dynamical analogs of chaos. To understand the two reasons, it is not necessary to review all existing definitions of structural stability. Rather, it suffices to describe their common logical form [Pugh and Peixoto, 2008]. Namely, given some equivalence relation R (e.g., topological conjugacy) over functions, one says a function f is structurally stable if all "close" functions (under some metric) are R-equivalent to f. Why are definitions of this form not analogous to characterizations of chaos (simpliciter)?

First, the concepts employed to define structural stability are disjoint from those used to define chaos. For example, definitions of structural stability typically use the notions of homeomorphism and diffeomorphism, whereas definitions of chaos employ notions like sensitivity to initial conditions, topological transitivity, density, etc. Of course, some difference in definitions is unavoidable, as structural stability is about small changes in time-evolution functions, whereas chaos is about small changes in states.

Nonetheless, if Werndl [2009] and Berkovitz et al. [2006] are correct, then probability is a key concept in characterizing degrees of chaos. In contrast, none of the definitions of structural stability employ probability at all. This is surprising, given that probability (and in particular, probabilistic independence) is perhaps the most widely-employed tool used to characterize uncertainty, noise, and (expected) error. The fact that probability is not used in definitions of structural stability, therefore, raises serious questions about the importance of such definitions for discussions of prediction, control, and explanation.[11]

Second, time plays different roles in definitions of chaos and structural stability respectively. Definitions of chaos typically contain a clause – like the definition of topological mixing – that places constraints on the distant future of the system. For example, in many chaotic systems, nearby initial conditions may have similar trajectories for a long period of time, but their trajectories may diverge suddenly and radically in the distant future. The potential for such sudden divergence is what renders long-term predictions problematic. In contrast, to my knowledge, all but one of the equivalence relations used to define structural stability constrain *only one time step* in the evolution of a dynamical system, and the one exception is typically only

[11] The reader will note that my definition of structural mixing likewise does not employ the use of probability. It turns out that the standard notion of topological mixing is closely related the ergodic (and hence, probabilistic) concept of strong mixing. I conjecture an analogous relationship will hold in the structural case, but this remains to be shown.

applied to dynamical systems that are described by differential equations.

These two reasons do not provide conclusive evidence that the mathematically rich research on structural stability is, at the end of the day, unimportant for empirical science. Rather, they suggest two more questions to add to the list at the outset of the paper: what are the relationships among various definitions of chaos and structural stability? And what is the importance of the various notions of structural stability for prediction, control, and explanation?

4 Conclusions and Philosophical Upshots

Section one outlined a broad research program, which consisted of three questions. Section two provided a brief example of how one might go about answering two of three questions. In particular, I defined a notion of "structural mixing" that is analogous to the standard notion of "topological mixing", and I proved a theorem relating the two concepts. I now conclude by discussing philosophical significance of this research program.

To see why this seemingly technical series of questions has broad philosophical importance, replace every occurrence of the phrase "time-evolution function" with the word "regularity" in the above discussion of structural chaos and in the two theorems. Doing so reveals that the main result roughly asserts that there are many "similar" regularities that (i) produce widely different future behavior and (ii) are compatible with the observed past. That's just an instance of the problem of induction. So investigating structural chaos amounts to investigating (in a mathematically precise setting) a (the?) central problem of epistemology and philosophy of science.

With this in mind, it is now easy to see why answers to each of the three questions are philosophically important. Question one asks, "For each definition of "chaotic system", what is the analogous concept of structural chaos?" Because there are different "degrees" of chaos [Berkovitz et al., 2006], an answer to question one would characterize differing "degrees" of problem of induction.[12] That is, an answer to the first question would allow one to characterize inductive problems in terms of their difficulty.

Question two asks, "what are the relationships among the various notions of chaos (simpliciter) and the analogous notions of structural chaos?" To see

[12]Kelly [1996] contains a sophisticated description of a hierarchy of "problems" of induction. I am skeptical there is any relationship between Kelly's hierarchy and that which would arise from pursuing the first question here. So this project would provide an independent, orthogonal way of characterizing inductive difficulty.

why this question is important, it is useful to consider one reason why chaotic systems are so interesting. The classic problem of induction shows that past observations are insufficient to identify a dynamical system's time-evolution function, and hence, there are many regularities that (a) are compatible with past observations and (b) predict radically different futures. The existence of chaos entails that predicting or manipulating a dynamical system's behavior might be difficult *even if the exact dynamics of the system are known.* Hence, an answer to question two provides a bridge between research on the classical problem of induction and new research in chaos theory, which respectively identify different sources of difficulty for prediction and manipulation.

Finally, question three asks, "what are the implications of structural chaos for prediction, control, and explanation?" The importance of this question is self-explanatory: prediction, control, and explanation are three central goals of science, and so an answer to question three amounts to an answer to the question, "Why is structural chaos important?"

A Proofs

Lemma 1 *Let X be any metric space, $U \subseteq X$ an open set and $F \subseteq X$ be finite. Then $U \setminus F$ is open. If X be has no isolated points, $U \setminus F$ is non-empty.*

Theorem 2 *Suppose φ is continuous and topologically mixing. Suppose that $\varphi \in \Phi$ and that Φ is closed under F-indistinguishability for some finite $F \subseteq X$. If X has no isolated points, then Φ is structurally mixing at φ.*

Proof: Let $x_0 \in X$. It must be shown that for all $\epsilon > 0$ and all non-empty open sets $V \subseteq X$, there is some $N \in \mathbb{N}$ such that

$$f_{x_0,n}(B_\epsilon(\varphi) \cap \Phi) \cap V \neq \emptyset \text{ for all } n \geq N$$

Call this condition $\dagger(\epsilon, V, N)$. Let $\epsilon > 0$ and $V \subseteq X$ be an open set.

Define $x_j = \varphi^j(x_0)$ for all natural numbers j, and let $M = |F| + 1$. Because Φ is closed under F-indistiguishability, there is $\beta > 0$ such that if (a) φ and ψ agree everywhere on all but finitely many elements of $X \setminus F$ and (b) $D(\varphi, \psi) < \beta$, then $\psi \in \Phi$. As φ is continuous and F is finite, it follows that for all $k \leq M$ there is $\delta_k > 0$ such that

$$B_{\delta_k}(x_k) \cap F = \begin{cases} \{x_k\} & \text{if } x_k \in F \\ \emptyset & \text{otherwise.} \end{cases}$$

and

$$y \in B_{\delta_k}(x_k) \Rightarrow d(\varphi(y), \varphi(x_k)) < \{\epsilon, \beta\}$$

Note here I am using $B_\gamma(z)$ to refer to the γ-ball around $z \in X$ with respect to the metric d, in the same way that I have used $B_\gamma(\varphi)$ to refer to the γ-ball around φ with respect to D.

Let $\delta = \min\{\delta_k : k \leq M\}$. Because φ is topologically mixing, for each $k \leq M$ there is $N_k \in \mathbb{N}$ such that for all $n \geq N_k$:

$$\varphi^n(B_\delta(x_k)) \cap V \neq \emptyset$$

Let $N_* = M + \max\{N_k : k \leq M\}$. I claim that $\dagger(\epsilon, V, N_*)$. Let $n \geq N_*$. It is necessary to find a function $\psi \in B_\epsilon(\varphi) \cap \Phi$ such that $\psi^n(x_0) \in V$. If $\varphi^n(x_0) \in V$, then we're done. So assume $\varphi^n(x_0) \notin V$.

Because $M > |F|$, there is $k \leq M$ such that $x_k \notin F$. Notice

$$n - k \geq N_* - M \geq \max\{N_j : j \leq M\} \geq N_k.$$

Hence, by choices of δ and N_*, there is $y \in B_\delta(x_k)$ such that $\varphi^{n-k}(y) \in V$. Note $y \neq x_k$ because $\varphi^{n-k}(x_k) = \varphi^n(x_0) \notin V$. I claim that y may be chosen so that $\varphi^j(y) \neq x_k$ for all $j \leq n - k$.

Why? Suppose for the sake of contradiction that for all $y \in B_\delta(x_k)$, there is some $j \leq (n-k)$ such that $\varphi^j(y) = x_k$. In particular, there is $j_0 \leq (n-k)$ such that $\varphi^{j_0}(x_k) = x_k$. Thus, for all $m \geq (n-k)$ and all $y \in B_\delta(x_k)$:

$$\varphi^m(y) \in \{x_k, \varphi(x_k), \ldots, \varphi^{j_0-1}(x_k)\}.$$

Let $T = X \setminus \{x_k, \varphi(x_k), \ldots, \varphi^{j_0-1}(x_k)\}$. Then T is non-empty and open by the lemma. However, by the above reasoning, $\varphi^m(B_\delta(x_k)) \cap T = \emptyset$ for all $m \geq (n-k)$. So φ is not topologically mixing, contradicting assumption.

It has been shown that $y \in B_\delta(x_k)$ may be chosen so that $\varphi^j(y) \neq x_k$ for all $j \leq (n-k)$. Define $\psi : X \to X$ as follows:

$$\psi(z) = \begin{cases} \varphi(y) \text{ if } z = x_k \\ \varphi(z) \text{ otherwise.} \end{cases}$$

Note $D(\varphi, \psi) = d(\varphi(x_k), \varphi(y))$. By continuity of φ, it follows that $d(\varphi(x_k), \varphi(y)) \leq \min\{\beta, \epsilon\}$. Hence, $\psi \in B_\epsilon(\varphi)$. Because ψ is equal to φ everywhere except $x_k \notin F$, it follows that ψ is βF-indistinguishable from φ. As Φ is closed under βF-indistinguishability, $\psi \in \Phi$.

Finally, $\psi^n(x) = \varphi^{n-k}(y) \in V$ because $\varphi^j(y) \neq x_k$ for all $0 \leq j \leq n - k$.

□

References

Robert W. Batterman. Defining chaos. *Philosophy of Science*, 60(1):43—66, March 1993.

Joseph Berkovitz, Roman Frigg, and Fred Kronz. The ergodic hierarchy, randomness and hamiltonian chaos. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 37 (4):661–691, 2006.

Seamus Bradley. Scientific uncertainty: A user's guide. *Grantham Institute on Climate Change*, Discussion Paper 56, 2012. URL `http://philpapers.org/rec/BRASUA`.

Robert L. Devaney, Luke Devaney, and Luke Devaney. *An introduction to chaotic dynamical systems*, volume 6. Addison-Wesley Reading, 1989.

Roman Frigg, Seamus Bradley, Hailiang Du, and Leonard A. Smith. Laplaces demon and the adventures of his apprentices. *Philosophy of Science*, 81(1):31—59, 2014.

Kevin T Kelly. *The logic of reliable inquiry*. Oxford University Press, New York, 1996.

Wendy S. Parker. Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 41(3):263–272, 2010.

Wendy S. Parker. When climate models agree: The significance of robust model predictions. *Philosophy of Science*, 78(4):579–600, 2011.

Charles Pugh and Mauricio Peixoto. Structural stability. *Scholarpedia*, 3(9), 2008. ISSN 1941-6016. doi: 10.4249/scholarpedia.4008. URL `http://www.scholarpedia.org/article/Structural_stability`.

Charlotte Werndl. What are the new implications of chaos for unpredictability? *The British Journal for the Philosophy of Science*, 60(1):195–220, 2009.

Contrastive explanations, crystal balls and the inadmissibility of historical information

Abstract

I argue for the falsity of what I call the "Admissibility of Historical Information Thesis" (AHIT). According to the AHIT propositions that describe past events are always admissible with respect to propositions that describe future events. I first demonstrate that this demand has some counter-intuitive implications and then argue that the source of the counter-intuitiveness is a wrong understanding of the concept of chance. I also discuss the relation between the failure of the AHIT and the existence of contrastive explanations for chancy events (which David Lewis denied).

Introduction

Suppose you know the chance of some event, E, and suppose this chance is very low. Then E occurs. Intuitively, this calls for an explanation. What *intuitively* calls for an explanation is not that E occurred. Rather it is that E occurred *rather than "not E"* (as "not E" had a greater chance of occurring). David Lewis famously argued that there can be no such explanation. There might be an explanation for E, but there is no contrastive explanation for "E rather than 'not E'". There is – argued Lewis – no reason for the outcome of a chancy event to turn out one way rather than another "for is it not the very essence of chance that one thing may happens rather than another for no reason whatsoever?" (Lewis 1986, p.175)[1]

At the macro-level, however, we often do give contrastive explanations for events that we also plausibly take to be chancy. For example, that I won the backgammon game I played yesterday rather than my opponent is explained by the fact that I am a much more experienced player (or by the fact that he was not paying full attention to the game, or by the fact that I was very determined to win so I spent a lot of time thinking before each move etc.). This is true, even though there was (up to the last turn of the game) a non-trivial chance that I will lose.

[1] See Percival 2000 for a good discussion of Lewis' position.

There were several (quite successful) attempts in the literature to give an account for such contrastive explanations that will be compatible with Lewis' claim regarding "the essence of chance". Here, however, I want to argue against Lewis' claim that it is the essence of chance that "one thing may happen rather than another for no reason whatsoever".

Notice that Lewis referred to "reasons" rather than to "causes". This is no accident. Reasons (unlike causes) justify beliefs. To say that there is no reason that A occurred rather than not A is to say that there is nothing that can justify a belief that A will occur rather than not A, over and above the fact that there was some chance that A rather than not A will occur.

Moving from full beliefs to partial beliefs, to say that there is no reason that A occurred rather than not A is to say that there is no propositions, E, such that one's degree of belief in A conditional on the proposition that says that the chance of A is x and E should be higher than one's degree of belief in A conditional on the proposition that says that the chance of A is x. If there is such a proposition then this proposition is a reason that A rather than not A will occur over and above the fact that there is some positive chance that A rather than not A will occur.

Although in his discussion of explanations of chancy events, Lewis does not explicitly commit himself to such a formulation, his choice of words clearly hints that this is what he had in mind, as the condition mentioned in the previous paragraph is a condition Lewis does explicitly discuss and endorse elsewhere (in Lewis 1980).

As it is, the condition is false and Lewis was well aware of that. A itself, for example, is a reason for the occurrence of A rather than not A. Lewis called propositions that describe such reasons, propositions that give information about the outcomes of chancy events over and above the information one gets by learning the chance of these events, inadmissible propositions.

Lewis was well aware that there are inadmissible propositions, but he believed many propositions are admissible. I take it that what Lewis really wanted to say in his discussion of contrastive explanations is the following: there is no propositions, E which is only about events prior to some time, t, before the occurrence of A, such that one's degree of belief in A

conditional on the proposition that says that the chance of A, at t, is x *and* *E* should be higher than one's degree of belief in A conditional on the proposition that says that the chance of A at t is x. In other words, Lewis wanted to say that there is no historical reason that A will occur rather than not A.

Lewis was explicitly committed to this latter claim. However, I will argue here, he was wrong. Past events can give us information about the occurrence of chancy future events, over and above the information we get by learning the chance of these future events. In the literature such propositions are sometimes called "crystal balls" (see for example Hall 1994). Although the term is catchy and successfully captures one aspect of the role they play – if they exist – in our systems of beliefs, it misses another important role. Balls made of crystal that show future events are very good in predicting these events, but they do not supply us (or the magicians that use them) explanations for the events they show.

Inadmissible propositions that describe past events, on the contrary, often do give us information about the future through the explanations they provide to future events (in case they will occur), or so I will argue. If this is so, then contrastive explanations are possible: a proposition E can serve as a contrastive explanation for another proposition, A, if E is inadmissible to A and is about events prior to the event described by A.

The rest of the paper will be organized in the following way. In sections 1 I will discuss Lewis' Principal Principle (PP) and the role the concept of admissibility plays in it. The discussion, I believe, will touch upon several issues that have not been properly dealt with in the literature. In section 2 I will discuss the claim that historical information is always admissible (call this claim the "Admissibility of Historical Information Thesis" or the AHIT). The main point of this section will be that the motivation for accepting the AHIT is that it enables the PP to perform the role it is supposed to play, i.e. to characterize the conceptual role of chance.

In section 3 I will argue that there are cases in which the AHIT does not intuitively play this role as it is inconsistent with another intuitive principle. In section 4 I will argue that Lewis' own theory of chance (which is designed to explain the PP) does not only allow but also predicts the

failure of the AHIT. In section 5 I will use the conclusions of the first four sections in order to defend Callender and Cohen (2010) and Hoefer (2007) from a recent criticism by Christopher Meacham (forthcoming).

The Principal Principle and the concept of admissibility

David Lewis was not the first to introduce the idea that a rational agent's degree of belief in a proposition, A, should be constrained by his beliefs regarding the chance of A. Long before Lewis published his 1980 paper in which he presented his version of the principle, the idea was well discussed in the literature under different titles ("The Principle of Direct Probability", "The Principle of Direct Inference", "Miller's Principle", "Probability Coordination". See Strevens [1999] for an overview).

Lewis' formulation of the idea has, however, several significant advantages over the formulations preceding it. One of these advantages is of special importance for the current discussion. In order to appreciate it, it will be instructive to first present what seems to be the most straightforward way to express the idea. Let us call it "the Naive Principle":

NP (naive principle): "A rational agent's credence in A, conditional on the proposition "the chance of A is x", equals x".

One problem with the NP is as follows. The principle is supposed to be a principle of rationality. It restricts the range of credence functions that a rational agent is permitted to adopt. However, if an agent starts with a rational credence function and then updates his beliefs after gaining new information in a rational way, he should end up holding another rational credence function. This is just part of what makes an updating method rational – that it preserves the rationality of credence distributions. The naive principle, however, is not necessarily preserved under any reasonable updating method, as after learning A the credence a rational agent assigns to A conditional on any other proposition must be 1, not the chance that A is true. Thus, it must be the case that credence distributions that do not obey the NP can be rational, which, in turn, means that the NP is not a principle of rationality.

Partly in order to handle this problem, Lewis introduced a variation of the naive principle that is not vulnerable to the problem just described. Lewis'

first formulation of the principle, which he called "the Principal Principle (PP), is as follows:

> Let C be any reasonable initial credence function. Let t be any time. Let x be any real number in the unit interval. Let X be the proposition that the chance, at time t, of A's holding equals x. Let E be any proposition compatible with X that is admissible, at time t. Then C(A|XE) = x. (Lewis 1980, p.266).

It is easy to see that the PP, unlike the naive principle, is preserved under Bayesian updating on A: it holds also after learning A because any *reasonable initial* credence function gives credence of 1 to A conditional on any proposition of the form "A and the chance of A is x". In other words, a proposition is always inadmissible to itself. Is the PP always preserved under Bayesian conditionalization? To see that it is, consider the following inference:

Let c(.) be the agent's initial probability distribution and let c'(.) be his probability distribution after learning some admissible proposition, E. Assume c(.) obeys the PP. Then:

$c'(A|XE) = c'(A|X) = c'(AX)/c'(X) = c(AX|E)/\ c(X|E) = c(A|XE) = x = c(A|X)$

Notice, that in order for the inference to be true no explication of the concept of admissibility is required. In order for Lewis' attempt to avoid the problem that the NP suffers from to work, it only has to be the case that

*For every admissible proposition, E, $c(A|\ XE) = c(A|X)$.

The plausibility of the PP depends, then, entirely on our willingness to accept - given an explication for "admissibility" - that * keeps on holding after Bayesian conditionalization on an admissible proposition.

Although * must hold in order for the PP to avoid the problem the NP suffers from, * cannot serve as a definition for admissibility (i.e. it cannot be the case that E is admissible to A iff * holds), as by defining admissibility in such a way it becomes impossible to violate the PP[2]. Thus, in order for the PP to have a bite, in order for it to restrict the range of credence functions that a rational agent is permitted to adopt, the concept of admissibility must be defined independently of the PP.

It is important to understand what exactly * demands, however. Let c''(.) be the agent's credence function after learning X. Then:

$c''(A|XE) = c''(A|E) = c''(AE)/c''(E) = c(AE|X)/c(E|X) = c(A|XE) = c(A|X) = c''(A)$

so

$c''(A|E) = c''(A)$

In other words, if E is admissible to A, then after learning the chance of A, E and A become probabilistically independent (even if prior to learning the chance of A, E and A were probabilistically dependent).

Indeed, Lewis understood admissibility exactly in this spirit:

> Admissible propositions are the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes. Once the chances are given outright, conditionally or unconditionally, evidence bearing on them no longer matters. (Lewis, 1980, p.272).

[2] Lewis was well aware of this point. He wrote: "The power of the Principal Principle depends entirely on how much is admissible. If nothing is admissible it is vacuous. If everything is admissible it is inconsistent" (Lewis, 1980, p. 272).

The above discussion makes it clear that admissibility is a triadic relation: it is a relation of one proposition, A, to another proposition, B, with respect to a given credence distribution, c(.). A proposition can be admissible to one proposition and inadmissible to another (every proposition, for example, is inadmissible to itself and admissible to any other proposition which is probabilistically independent of it), and a proposition can be admissible to another proposition with respect to one credence distribution, but inadmissible to it with respect to another (for example, if I believe to degree 1 that every time I flip a coin using my left hand, it falls "Heads" , then "I flipped the coin using my left hand" is inadmissible to "the coin falls Heads", but if I believe this conditional to degree 0, then the admissibility relation between the two propositions does hold).

Which propositions are admissible?

We saw in the previous section that both the power and the plausibility of the PP depend on how much is admissible. Lewis' informal characterization of admissible propositions (that was quoted above) captures the role the concept plays in rational reasoning. However, it does not help one determine whether a given proposition is admissible.

Lewis did, however, characterized two families of propositions that must be admissible. The first family is that of propositions about the past: At any given point in time, t_i, every proposition which is only about events prior to t_i, is admissible to any proposition, E, which is about future (relative to t_i) events.

The second family is that of conditionals in which the antecedent is a complete description of the world up to some point in time, t_i, and the consequent is a proposition that assigns a certain chance to some event, E, at t_i (Lewis added one qualification for this characterization, but it should not concern us here). All such propositions, argued Lewis are admissible to E (at all times).

Using these two claims Lewis introduced a second version of the PP and showed that it follows (using his two assumptions) from the first version. Here it is:

Let H_t be a complete description of the world up to time t; let T be a conjunction of conditionals of the sort just described (i.e. conditionals from full histories of the world up to a time, t, to chances of events at t) that assigns a chance to every event at t; let $P_t(.)$ be the chance distribution over a set of events according to T at time t; let c(.) be any reasonable initial credence function and let A be any proposition to which T assigns a chance at t, then:

$$c(A|TH_t) = p_t(A)$$

While the second version of the PP follows from the original version, it is not clear (without a full characterization of admissibly) whether the two versions are equivalent[3].

Christopher Meacham (2010) argued that Lewis intended the two versions to be equivalent and suggested to take their equivalence as a criterion for admissibility: He introduced a formal definition for admissibility and proved that this definition is necessary and sufficient for the two versions to be equivalent.

There is no need for us to discuss Meacham's condition. Given Meacham's criterion for admissibility – namely that it must make the two versions of the PP equivalent – his condition is the right one to adopt. The problem is that Meacham adopted the wrong criterion. It is the wrong criterion *because* it makes the two versions of the PP equivalent. The two versions cannot be equivalent, I will argue in the next section, because while the first version is a principle of rationality, the second is not.

The problem with the second version to which I will point is its commitment to the claim that past events must be admissible to any future event. Let us call this commitment the Admissibly of Historical Information Thesis (AHIT). Before arguing against the AHIT, it will be instructive to explain the initial motivation for accepting it.

[3] In this paper I do not discuss Lewis' "new principal principle" or any of the other attempts to solve "the big bad bug" as none of these attempts is relevant to my argument.

Lewis does not explicitly discuss his reasons for adopting the AHIT. Moreover, he does explicitly claim (see Lewis 1980 p. 274) that the AHIT is only true "as a rule" and might have rare exceptions. He also claims that it being true "as a rule" is a contingent matter that might be absent in other possible worlds. His reasons for these qualifications of the AHIT are the following. Lewis pointed to the possibility of what was later described by Ned Hall (1994) and others as "crystal balls", i.e. past events that carry information about the future outcomes of chancy events:

> "if the past contains seers with foreknowledge of what chance will bring, or time travelers who have witnessed the outcome of coin tosses to come, then patches of the past are enough tainted with futurity so that historical information about them may well seem inadmissible" (Lewis 1980 p. 274).

Meacham (2010) presented an argument against the possibility of crystal balls. I will critically discuss his argument in section 5 and argue that there are in fact many crystal balls in our world. We all know them: they are described by the "special sciences".

In any case, it seems that in the absence of crystal balls, Lewis would be willing to accept the AHIT as always true (and as noted, Hall, Meacham and others explicitly do so). Why did he find the AHIT so attractive?

The reason, I believe, does not steam from Lewis' commitment to a specific theory of chance. Rather it lays in the conceptual role Lewis took the PP to play. Lewis took the PP to express all "that we know about chance" (Lewis, 1980. P. 266). Whatever chance is, Lewis believed, it must make the PP a principle of rationality. Our concept of chance, according to Lewis, is a concept of a feature of reality that plays the role the PP assigns to chance. Indeed, for Lewis, a restriction on any theory of chance is that it must explain why the PP is a principle of rationality (see his discussion in Lewis 1994).

Lewis' acceptance of the AHIT should, I believe, be understood in a similar way. It is clear that without an operational characterization of a large family of propositions which are admissible, there is no conceptual role that can be ascribed to chance via the PP. To say that the PP does capture the conceptual role chance plays is to commit oneself to some such a characterization (and, as was explained in the previous section, some propositions must be taken to be inadmissible in order for the PP to be consistent). Lewis thought that in our world this characterization is partly captured by the AHIT.

This is, I believe, why Lewis argued (in the quote above) that the existence of crystal balls might make propositions about the past "seem inadmissible". It is clear from his discussion that he did not only take historical information in worlds in which there are crystal balls to *seem* inadmissible but that he also took them to actually *be* inadmissible: they are inadmissible *because, in such worlds, they seem inadmissible*.

In such worlds, the PP captures the conceptual role chance plays using a different set of propositions which are always admissible, not using the set of all propositions that carry only historical information.

However, in our world – Lewis' thought was – the PP does capture the conceptual role chance play using the AHIT. Lewis (implicitly) supported this claim using an example of a coin which you are certain is fair (i.e. you are certain that its chance to fall Heads is 0.5). Lewis wrote:

> "...you have plenty of seemingly relevant evidence tending to lead you to expect that the coin will fall heads. This coin is known to have a displaced center of mass, it has been tossed 100 times before with 86 heads, and many duplicates of it have been tossed thousands of times with about 90% heads. Yet you remain quite sure, despite all this evidence, that the chance of heads this time is 50%. To what degree should you believe the proposition that the coin falls heads this time?"
>
> Answer. Still 50%. Such evidence is relevant to the outcome by way of its relevance to the proposition that the chance of heads is 50%, not in any other way. If the evidence somehow fails to diminish your certainty that the coin is fair, then it should have no effect on the

> distribution of credence about outcomes that accords with that certainty about chance. To the extent that uncertainty about outcomes is based on certainty about their chances, it is a stable, resilient sort of uncertainty-new evidence won't get rid of it." (Lewis 1980, pp. 265-6).

Notice that Lewis does not present in this quote an argument for the application of the AHIT he describes. He just states it. This is so; it should be clear by now, since Lewis took this example to be a paradigmatic example of our everyday use of the concept of chance. To the extent that some feature of reality plays the role of chance in this example, the application of the AHIT Lewis employs must be valid on conceptual grounds.

I do not disagree with Lewis about his use of the AHIT in this example. I do, however, believe that not all applications of the AHIT are equally self-evident.

In order to convincingly argue against Lewis that in our world the AHIT is sometimes false (i.e. that there are crystal balls in our world) I must, then, demonstrate that in our world there are instances of historical information which is *intuitively* inadmissible. I will do this now.

Inadmissible historical information

Let E be a proposition that describes some future event. Let E* be a proposition of the form "the chance of E at time t is x". Let M be a proposition that describes some past event (i.e. an event prior to t) that *intuitively explains* E. Let c(.) be the credence function of a rational agent at time t.

> Epistemic relevancy of explanations (ERE): there are some cases in which

$$c(M|E^*) < c(M|E^*E)$$

ERE demands that in some cases learning that some event, E, has occurred makes a rational agent raises the credence he attaches to some past event, M, occurring (i.e. to an explanation of E), even if the agent already learned what the chance of E was just before its occurrence (i.e. at a time later than the time in which M has occurred).

Example 1: the conditional credence I attach to "Bob intended – a second ago - to open the window" given that Bob will actually open the window is higher than the unconditional credence I attach to this proposition and this is so even if I already learned the chance of Bob opening the window (now). In other words, Bob actually opening the window is more indicative of Bob's intention to open the window than the proposition that there is some positive – but possibly very low – chance of Bob opening the window (the chance of Bob opening the window can be low, for example, either because Bob did not intend to open the window and there is a low chance that Bob will do it involuntarily, or because Bob did intend to open the window but there is a high chance that Bill the Bully will not let him open the window. Learning that Bob actually opened the window clearly indicates that there was an intention on Bob's behalf).

The example demonstrates that there are cases in which giving up on the ERE (i.e. denying the claim that the epistemic state described in the example is rational) is extremely unintuitive. However, it is straightforward to see that the ERE is inconsistent with the AHIT.

Proof: $c(M|E^*E) = \frac{c(ME^*E)}{c(E^*E)} = \frac{c(E|E^*M)c(M|E^*)c(E^*)}{c(E^*E)} = \frac{c(E|E^*M)c(M|E^*)}{c(E|E^*)}$

However, if the AHIT holds and c(E|E) = c(E|E*M)*

$\frac{c(E|E*M)c(M|E*)}{c(E|E*)}$ = *c(M|E*).*

Thus,

*c(M|E*E) = c(M|E*) contrary to the ERE.*

In words: if learning that Bob opened the window is more indicative to his intention to open the window than that there is a certain (high or low) chance that he will open the window, then Bob's intending to open the window is inadmissible to Bob actually opening the window.

It is clear, then, that either the ERE or the AHIT must go. Now, here is what I would like to argue: since the original motivation for adopting the AHIT is the intuitive appeal of its role in a characterization of the conceptual status of the concept of chance and since in the example the AHIT intuitively does not play this role, the conclusion must be that it is the AHIT that has to go[4].

I do argue that but arguing only that is not enough. The situation is a bit trickier here. When I look at the ERE applied to our example my intuition strongly supports accepting it. However, I must admit that when I look at the AHIT applied to our example my intuition also support accepting it.

My degree of belief that Bob intended to open the window given that there was a chance of 0.01 that he will does intuitively seem to me to be much lower than my degree of belief that Bob intended to open the window given that there was a chance of 0.01 that he will and that he will actually open the window. But, at the same time, I find the demand to fix my degree of belief that Bob will open the window given that there is a chance

[4] It is important to emphasise that the example only aims at showing that in some cases the AHIT is unintuitive, not that it is false. However, as explained, the only justification for the AHIT is its intuitiveness as a restriction on the PP that helps it perform its role as a characterization of the conceptual role of chance.

of 0.01 that he will and that he intended a second ago to open the window on 0.01 highly intuitive.

My intuitions conflicts. Now, I have argued that since the ERE must be true (based on the example), the AHIT cannot be true. However, one can make the opposite inference. One can argue that since the AHIT must be true, the ERE cannot be true[5]. Both alternatives – rejecting the AHIT and endorsing the ERE and vice versa - are consistent.

In order to convincingly argue for rejecting the AHIT while endorsing the ERE it is not enough for me to point to how unintuitive rejecting the ERE is. I must also explain why the intuition that supports accepting the AHIT is misleading. Part of my explanation will be presented only in the next section, but the general strategy can be presented now.

My explanation begins with a diagnosis for why we find the AHIT so intuitive. Here it is: we tend to think about chance as something like “degrees of belief of a perfectly rational and maximally informed agent”. Since such an agent will not ignore relevant information such as “Bob intended a second ago to open the window” when forming his beliefs about the prospects of Bob opening the window, we let the chance “screen off” the effect of learning that Bob intended a second ago to open the window.

This way of thinking about chance is actually explicitly adopted by Ned Hall (1994 and 2004) and others (see footnote 7 in Hall 1994 for example). Hall writes: "Why should chance guide credence? Because-as far as its epistemic role is concerned-chance is like an expert in whose opinions about the world we have complete confidence. Let us imagine that chance just is the credence of such an expert, called "the Oracle". Since the Oracle's credence is ideal, we should like our own to match hers "(Hall 1994, p. 551).

I tend to agree that *if* Hall is right and chance is like an expert in whose opinions about the world we have complete confidence, the AHIT should hold. I do not, however, agree with Hall that chance is always like such an expert. In the next section I will argue that although chance must be like

[5] I thank Matthew Cotzen for this formulation of the dilemma.

such an expert when it comes to some types of propositions, chance cannot be like such an expert regarding all types of propositions.

Thus, although the AHIT does seem intuitive, in some cases this intuition is misleading and the reason for that is that it is based on a wrong understanding of the concept of chance.

Before presenting my argument for this last claim, Notice that the inadmissible proposition in the example (that Bob intended to open the window) can serve as an explanation for the proposition it is inadmissible for (that Bob will actually open the window). In fact, prior to learning whether Bob will open the window (but possibly after learning the chance that Bob will open the window) the agent *can be certain* that Bob's intention to open the window will serve as an explanation for Bob opening the window, in case Bob will open the window. The agent is just uncertain whether there will be anything to explain (i.e. whether Bob will open the window).

More importantly, if (as I argued) Bob's intention to open the window is inadmissible to "Bob will open the window" then Bob's intention can serve as a *contrastive explanation* for why Bob opened the window *rather than not*. Suppose I learned that the chance of Bob opening the window is very low. Then Bob opens the window. This, of course, is a surprise so it is natural to ask: why did Bob open the window *even though there was a very low chance that he will*? "because he intended to" seems like a good answer to this question.

This is so, however, only under the assumption that Bob's intention is inadmissible to Bob's act. If Bob's intention is admissible to Bob's act then it cannot teach me anything about Bob's act above what it teaches me about the chance of Bob acting in a certain way. Thus, the answer "because he intended to" does not seem appropriate: Bob's intention does not explain why Bob opened the window rather than not *even though there was a low chance that he will*, because Bob's intention contains information about Bob's act only through the information it contains about the chance of Bob's act.

However, inadmissible information does teach a rational agent something about a proposition not through what it teaches him about the chance of

this proposition. Thus, if Bob's intention is inadmissible to Bob's act, it can explain why Bob opened the window even though there was a low chance that he will. Since intuitively it does explain that, intuitively it is inadmissible to "Bob will open the window".

Now, notice that Bob's intention – if it is, as argued, inadmissible to Bob's action - is a kind of a crystal ball: it is an event that contains information about the future outcome of a chancy event. However, this information seems intuitively to be exactly the kind of information needed in order to explain the outcome of the chancy event, in case it will occur.

To better demonstrate this, consider the two following examples (structurally identical to example 1):

Example 2:

My conditional degree of belief, at 9:01, in "there was a cloud in the sky a minute ago", given "it will rain at 9:10 and the chance – now – that it will rain at 9:10 is x" is higher than my conditional degree of belief in "there was a cloud in the sky a minute ago" given only "the chance that it will rain at 9:10 is x". This is intuitively so since under the assumption that it will actually rain at 9:10, it is very likely that there was a cloud in the sky a minute ago (even if that was a cloud with low chances of dropping rain). However, only under the assumption that there is a low chance for rain, it is unlikely that there was a cloud in the sky.

Example 3:
My conditional degree of belief in "Bob is a much more experiences Backgammon player than Ann" given "Bob will win the game and the chance – now – that Bob will win is x" equals my conditional degree of belief in "Bob is a much more experiences Backgammon player than Ann" given only "the chance that Bob will win the game is x".

While in example 3 the two conditional probabilities seem intuitively to be equal, in – the structurally equivalent – example 2 they seem intuitively to be different of each other. What is the difference between the examples?

Well, while in example 2 the proposition "there was a cloud in the sky a minute ago" intuitively explains the proposition "it will rain at 9:10" even under the assumption that at 9:01 there was a chance of x for it to rain at 9:10, in example 3 the proposition "Bob is a much more experienced player" does not explain "Bob will win" under the assumption that just before the game ended Bob had a chance of x to win the game.

Thus, it seems that a proposition's inadmissibility with respect to another proposition and its explanatory power with respect to that proposition are closely related.

In this section I argued for the un-intuitiveness of a mathematical implication of the AHIT. In the previous section I explained that such un-intuitiveness is problematic since the supposed justifications of the AHIT exactly is that it intuitively enables the PP to perform its role. Thus, together the conclusions of the two sections constitute an argument against the AHIT. However, I did not supply an explanation for what makes historical information inadmissible, in cases it is. As explained, supplying such an explanation will make my argument against the AHIT even stronger. In order to explain why historical information can be inadmissible we must turn to specific explications of “chance”. In the next section I will argue that adopting “best system” explications of chance (such as Lewis’ 1994 own theory or Carl Hoefer’s 2007 version) can explain the inadmissibility of historical information.

The best system explication of chance and inadmissible historical information

According to both Lewis and Hoefer the chance of an event is the chance the best chance system of the world attaches to this event. Although there are several important differences between Hoefer’s version and Lewis’ original one, they should not concern us at this point.

Instead of committing myself to a specific "best system" account and describing it in length, it might be better to demonstrate the way my explanation for the failure of the AHIT works using a simple example to which both accounts will give approximately the same treatment.

Consider a (very simple) world in which, at every point in time, the universe can be in one of two possible states, 1 or 0. Suppose in this world there are only 20 discrete points in time. Finally suppose this world can be described using the following finite sequence of zeros and ones.

00110000001110101100

In this world, at the first point in time event of type 0 occurs, at the second point in time event of type 0 occurs again, at the third point in time event of type 1 occurs and so on. Let us call events which are the state of the universe at each point in time, basic events.

The sequence above is, we stipulated, a full *description* of the world. A *chance system* of this world is a set of claims (that can be, of course, encoded as another sequence of 0s and 1s) of the following form: "at t_i the chance of event E is x" that obeys the Kolmogorov axioms[6]. A *full* chance system of the world is a chance system of the world such that for every t_i and for every event, E (not only basic events but also events which are unions and intersections of basic events), there is a x such that the claim "at t_i the chance of event E is x" follows from the system.

A full chance system of the world is trivial if all the chance values it assigns to events are either 0 or 1 and is *nontrivial* if the chance value it attaches to at least one event is strictly higher than 0 and lower than 1. Lewis only deals with nontrivial systems that assign at each point in time a trivial chance value to any event prior to that point in time and I will follow him in that (the last assumption - that the chance of any past event is trivial - is explicitly rejected by Hoefer 2007 but I do not need to relax this assumption in order to make my point).

[6] It is straightforward to see that this characterization of chance systems is identical to the one Lewis' used and that was mentioned in section 2: a conjunction of conditionals from complete histories up to a time to a chance value for an event after that time.

The *best* system of this world is, according to both Lewis and Hoefer, a system of this world that has the best balance between strength, simplicity and fit. Obviously, a lot hangs on the questions of what constitutes the right measures of strength, fit and simplicity and on what constitutes the right "balance" between the three. Neither Lewis nor Hoefer presented accurate answers to these questions[7]. However, I think in the context of our simple example it will be fair to say that both of them will take the "strength" of any system of the simple world we are considering to go roughly with "how close" the system is to the ideal of being a full system (i.e. of assigning a chance value to all events); "simplicity" goes roughly with the length of the system and "fit" goes roughly with "how close" the chance values the system assigns to different events are to the actual relative frequencies of these events.

In order to demonstrate how historical information can be inadmissible it will be convenient to restrict our attention only to full systems. This will allow us to examine more freely the trade-off between fitness and simplicity.

Obviously, the fittest full system of our world is the system that assigns, at each point in time, a chance of 1 to any event if it occurs and a chance 0 if it does not. This system is, however, not very simple. A simpler full system of our world will assign to some future events non-trivial chance values.

Consider first T_0, a very simple system of our world. T_0 assigns, at each point in time, a probability for a future basic event to be of type i (when i is either 0 or 1) which is equal to the actual relative frequency of basic events of type i in the sequence.

Let us use the notation $E_{t_j}^{\,i}$ for "the basic event that occurs at t_j is of type i" (when i is either 0 or 1). Let $P_{t_i}(.)$ be the chance function according to T_0 at

[7] Maybe except with regards to "fitness" in which more precision can be found. Lewis understood fitness in the following way: "the chance of that course of history will be higher according to some systems than according to others" (Lewis 1994, p.480). Elga (2004) argues that this is not a satisfactory way to understand fitness and suggested an alternative explication that Hoefer (2007) adopts. My treatment of fitness is consistent with both approaches, as I will only apply the concept to our very simple world in which the disagreements between the two explications of "fitness" are mute.

t_i (of course, the chance values T_0, or any other system, assigns to every event may change at different points in time).

At each point in time:

For any $y \geq 1$:

$P_{ti}(E_{ti+y}^{1}) = 2/5$

$P_{ti}(E_{ti+y}^{0}) = 3/5$

And for every $i,j \in \{0,1\}$

$P_{ti}(E_{ti+y+1}^{i} | E_{ti+y}^{j}) = P_{ti}(E_{ti+y+1}^{i})$

Here is the full description of our world again:

00110000001110101100

I leave it to the reader to verify that I have constructed T_0 correctly.

Although T_0 is a very simple system, it is not very fit. It might be the case that the best system of our simple world is a little less simple though a little fitter. Consider, for example, T_1 which is the system that we get by equating the probability of the next basic event to be of type i, at each point in time in which the basic event is of type j, to the actual relative frequency of "i"s after "j"s which are not the last event in the sequence*.*

(The addition in Italics in the previous sentence is necessary, since after the last event in the sequence comes no other event. Thus, we can look instead, on the relative frequency of "i"s after "j"s which are not the last event in the sequence).

In our case, this is what we get:

Let Q(.) be the chance function according to T_1.

For any y ≥ 1:

At each point in time in which the basic event is of type 1:

$Q_{ti}(E_{ti+1}^{0}) = 0.5 = P_{ti}(E_{ti+y+1}^{0} | E_{ti+y}^{1})$

$Q_{ti}(E_{ti+1}^{1}) = 0.5 = P_{ti}(E_{ti+y+1}^{1} | E_{ti+y}^{1})$

At each point in time in which the basic event is of type 0:

$Q_{ti}(E_{ti+1}{}^0) = 7/11 = P_{ti}(E_{ti+y+1}{}^0 | E_{ti+y}{}^0)$

$Q_{ti}(E_{ti+1}{}^1) = 4/11 = P_{ti}(E_{ti+y+1}{}^1 | E_{ti+y}{}^0)$

Here is the full description of our world again:

00110000001110101100

At any point in time in which the basic event is of type 1, $Q_{ti}(E_{ti+1}{}^0) = 0.5$ is true, for example, because in our sequence there are eight 1s (and none of them is the last event in the sequence) and after four of them there is another 1.

At any point in time in which the basic event is 0, $Q_{ti}(E_{ti+1}{}^0) = 7/11$ is true because in our sequence there are eleven 0s which are not the last event in the sequence (there are 12 0s overall, but one of them is the last event in the sequence) and after seven of them there is another 0.

The best system of our world might be fitter but less simple than T_1 (this depends on the exact way in which the balance between fitness and simplicity is constituted), but for our demonstration we can assume that T_1 is the best theory of our world, as the point I am trying to push can be made using any non-trivial system.

The point is, actually, very simple. Even in the very simple world we are considering there are many regularities not captured by T_1 (or any other non-trivial system). For example, after any sequence of the form 001, the next basic event is 1. Thus, although at the third point in time, for example, $Q_{t3}(E_{t4}{}^1) = 0.5$, an agent that assigns, at t_3, a credence of 1 to the event $E_{t4}{}^1$ will – others things being equal - get his credence values "closer to the truth" (regarding the basic event at t_4) than an agent who assigns at t_3 a credence of 0.5 to this event. Such an agent must take, at each point in time, t_i, events of the form $E_{ti-2}{}^0 \cap E_{ti-1}{}^0 \cap E_{ti}{}^1$ to be inadmissible to events

of the form E_{ti+1}^1, i.e. he must violate the AHIT[8]. What can be a plausible ground for arguing that such an agent is irrational?

The agent is not irrational in the sense of violating either the Kolmogorov axioms or the PP (i.e. the original version of the PP that includes the admissibility clause) and he does adopt a credence distribution which is "closer" to the actual relative frequencies in the world. It is, of course, true that the agent does not know that this is the case. Thus, one might argue, there seems to be no good reason for him to take events of the form $E_{ti-2}^0 \cap E_{ti-1}^0 \cap E_{ti}^1$ to be inadmissible to events of the form E_{ti+1}^1. This might be true but the important point is that the agent does not have any good reason *not* to take such events to be inadmissible to events of the form E_{ti+1}^1. Thus, it seems that rationality allows (but certainly does not dictate) treating events of the form $E_{ti-2}^0 \cap E_{ti-1}^0 \cap E_{ti}^1$ to be inadmissible to events of the form E_{ti+1}^1.

Furthermore, an agent might gain – through inductive inference - good reasons to take certain kind of (historical) events to be inadmissible to (future) events of another kind. To the extent that inductive inferences can give an agent good reasons to adopt a certain degree of belief in a given (either full of partial) chance system, it can also give the agent good reasons to treat a certain class of propositions as inadmissible to some other class of propositions (more on this point in the next section)[9].

It will be useful to explicitly state a possible flawed objection to my treatment of the example in this section. One might argue that it was wrong of me to assume that T_1 is the best system of the world since it must be the case that the best system of the world assigns at t_3 a probability of 1 to E_{t4}^1. In other words, the objection is that given that there is at least one proposition, A, which should (or could rationally) be taken to be inadmissible to another proposition, B, relative to some chance system,

[8] Explicating the term "closer to the truth" here might prove to be a tricky business, but it seems to me obvious that any plausible explication of the term must take a credence distribution (at t_3) that obeys the PP with respect to T_1 but takes events of the form $E_{ti-2}^0 \cap E_{ti-1}^0 \cap E_{ti}^1$ to be inadmissible to events of the form E_{ti+1}^1, to be closer to the truth than a credence distribution that obeys the PP with respect to T_1 but does not take these events to be inadmissible.

[9] See Smart (2013) for a good discussion of Humean vs. anti-Humean treatments of induction.

T_i, there is another system, T_j which is better that T_i. This system is, so the objection goes, the system that agrees with the credence distribution of an agent who takes T_i to be the best system of the world and treats A as inadmissible to B.

In order to see what is wrong with this objection, let us go back to our example and let us assume (with the objection) that the best system of the world, let us call it T_2, assigns to any event a chance value which is equal to the credence value a rational agent that takes T_1 to be the best chance system of the world but also takes events of the form $E_{ti\text{-}2}{}^0 \cap E_{ti\text{-}1}{}^0 \cap E_{ti}{}^1$ to be inadmissible to events of the form $E_{ti+1}{}^1$ would assign to it.

T_2 is, no doubt, fitter than T_1, but it is also less simple as it adds some qualifications to T_1. Now, it surely might be the case that the best system of the world *is* less simple (but fitter) than T_1, but the point is that the best system of the world is not the fittest system (i.e. the system that assigns at each point in time a chance of 1 to any event that occurs and a chance 0 to any event that does not). Thus, whatever the best system is, a rational agent can "do better" (in terms of fittness) than the best system by taking some historical events to be inadmissible to some future events.

To demonstrate, note that T_2 still misses some regularities in our simple world. Here is the description of our world again:

00110000001110101100

For example, at each point in time, t_i, in which it is true that $E_{ti\text{-}2}{}^0 \cap E_{ti\text{-}1}{}^0 \cap E_{ti}{}^1$ it is also true that $E_{ti+3}{}^0$. Thus, a rational agent can still "gain" in terms of fitness by setting, at t_5 and at t_{13},

$c(E_{ti+1}{}^0 | T_2 \cap E_{ti\text{-}2}{}^1 \cap E_{ti\text{-}3}{}^0 \cap E_{ti\text{-}4}{}^0) = 1$

i.e. by taking events of the form $E_{ti\text{-}2}{}^1 \cap E_{ti\text{-}3}{}^0 \cap E_{ti\text{-}4}{}^0$ to be inadmissible to events of the form $E_{ti+1}{}^0$, relative to T_2.

Notice that, unlike the case of events of the form $E_{ti\text{-}2}{}^0 \cap E_{ti\text{-}1}{}^0 \cap E_{ti}{}^1$ which can be rationally taken to be inadmissible to events of the form $E_{ti+1}{}^1$ relative to T_1, events of the form

$E_{ti-2}^{1} \cap E_{ti-3}^{0} \cap E_{ti-4}^{0}$ can be rationally taken to be inadmissible to events of the form E_{ti+1}^{0} relative to T_2, *even though the type of events T_2 was designed to be the fittest with regards to which do not supervene on them* (in the following sense: while events of the form $E_{ti-2}^{i} \cap E_{ti-1}^{j} \cap E_{ti}^{k}$ determine events of the form E_{ti}^{k}, events of the form $E_{ti-2}^{i} \cap E_{ti-3}^{j} \cap E_{ti-4}^{s}$ do not determine events of the form E_{ti}^{k}). Thus, my explanation for the failure of the AHIT does not hold only in cases in which the type of events the best system of the world is designed to deal with supervene on the type of inadmissible historical events which are responsible for the failure of the AHIT.

The explanation offered here for the failure of the AHIT is very simple: as long as the best system of the world in non-trivial, for any given level of simplicity, there are some regularities in the world that the best system does not capture and these regularities can justify taking some historical events to be inadmissible.

Although the explanation is very simple it sheds new light on the role the admissibility clause plays in the PP. Since a chance system must obey the laws of probability, holding the levels of simplicity and strength of the system constant, a chance system cannot be "the fittest" with respect to all types of events. It must "choose" the type of events regarding which it seeks to be the fittest[10]. Whatever this type of events is, there is another type of events regarding which there is another system which is fitter than the best system of the world.

The admissibility clause enables rational agents to overcome this limitation of chance systems. While the level of simplicity of a chance system has a constitutive role in what makes or does not make it the best system of the world, it plays no such role in what makes a given credence distribution a rational (or irrational) one. While the "goodness" of chance systems is sensitive to how simple they are, the rationality of credence functions is not. Simplicity is a theoretical virtue, it is not a virtue for rational agents. *A rational agent should never give up on the accuracy of his degrees of belief in order to make his credence function simpler*. He should, that is, use all the information available to him, even if this means

[10] Lewis believed, for example, that this type of events (the type of events with respect to which the fitness of the best theory is measured) is micro-physical events.

taking more events in the algebra over which his credence function is defined to be probabilistically dependent.

The admissibility clause enables rational agents to do just this. It allows them to use information that the best system of the world ignores for the sake of simplicity. This explains how historical information can be inadmissible, i.e. it explains how past events can teach a rational agent something about future events not through what they teach him about the chances of these future events. They can do that by giving the agent information about the future which is ignored by the best system and thus is not reflected in the chances of future events.

This also explains why it is wrong to understand chance (or chance's epistemic role) in the way Hall suggested (i.e. as an expert). An expert's credence function should always be based on all the information available. Thus, when it comes to (a maximally informed about the past) expert it makes sense to let the expert's opinions "screen-off" the evidential support of past propositions. However, unlike experts, chance values are constituted in a way which always ignores some information. When an agent has good reasons to suspect that he has stumbled upon such information, he is rationally permitted (and plausibly required) to use it.

This concludes my positive argument for the possibility of violations of the AHIT. In the next section I will argue against a possible objection to my account, presented by Meacham (forthcoming).

Meacham's argument against autonomous chances

The explanation suggested in the previous section for the failure of the AHIT, is closely related to a claim made my Hoefer in his 2007 paper. Hoefer argues in his paper that different physical set-ups may give rise to different chance values. Very roughly the idea is that as long as a given physical set up produces a sequence of events in which a stable regularity that is best characterize using a chance system is observed, it is justified to assign the chances – according to this characterization – the name "chance".

At the end of his paper Hoefer discusses a situation in which a rational agent knows the chance of some macro-level event (such as a train

arriving at the station at some time) according to both the best system of chance (which is constructed – this is Hoefer's assumption – to be the fittest system regarding micro-level events) and some high-order system that assigns a chance value to the macro-level event according to some regularity in the macro-level set-up (such that the set-up that consists of all the arrivals of the train to the station):

> Suppose God whispers in one ear the macro-level chance, based on the entire history of 9:37 trains in my town, while a Laplacean demon who calculates the micro-derived chance whispers it in your other ear. Which should you use? Common
>
> wisdom among philosophers of science suggests that it must be the micro-derived chance... But on the contrary, I want to suggest that it could be the macro-derived chance that better deserves to guide credence. How could this be? (Hoefer 2007, p.592).

Hoefer's answer to the question he poses is the following:

> The micro-level chances are what they are because they best systematize the patterns of outcomes of micro-level chance setups, such as quantum state transitions... But that entails nothing about what will happen for train arrivals. The micro-theory of chances in the Best System gets the frequencies right for micro-level
>
> events (and reasonable-sized conjunctions and disjunctions of them), to a good approximation, over the entire mosaic. This simply does not entail that the micro-theory must get the frequencies right for sets of distinct one-off setups, each being a horribly complex conjunction of micro events... (Hoefer 2007, p. 593).

Hoefer goes on to generalize this conclusion:

> There are chance-making patterns at various ontological levels. Nothing makes the
>
> patterns in one level automatically dominate over those at another; at whatever
>
> level, the chances that can best play the PP role are those that count as the 'real'
>
> objective probabilities. (Hoefer 2007, ibid)

It is easy to spot the similarities between Hoefer's argument and the one presented in the previous section. As the best system of the world, T_{best}, is designed to be the fittest regarding a given type of events, and as it must keep its level of simplicity relatively high, there will always be some other type of events regarding which the best system is inferior, in terms of fitness, to another chance system, $T_{alternative}$. A rational agent, it seems natural to argue, should use the chance system that best fit the type of events he is considering to guide his credence, not the best system overall.

It seems clear to me, however, that if this is right, then the conclusion of the previous section must hold: a rational agent must sometimes (i.e. when he assigns a relatively high credence to $T_{alternative}$) equate his conditional credence in some proposition, P, given that the chance of P (according to T_{best}) is x and some other proposition, E, (which is a proposition of the type regarding which $T_{alternative}$ is the fittest system) to the chance of P given E according to $T_{alternative}$, not to x. Thus, if all of this is correct, the PP can be taken to be a principle of rationality only if in such a case, E is taken to be inadmissible to P.

Hoefer does not explicitly adopt this conclusion in his paper. In a puzzling footnote he writes: "It may be that our two posited chances are such that admissibility considerations rule out the use of one, if the other is known, as we saw in the breast cancer case. *But it is not clear to me that this must happen in general*" (Hoefer 2007, footnote 35, my *italics*). I do not

see, though, how it is possible to reject my conclusion, while accepting Hoefer's conclusion and taking the PP as a principle of rationality.

In a recent paper, Meacham seems to agree with this last claim. He, however, takes it as an argument against the claim that sometimes it is $T_{alternative}$ that should guide a rational agent's credence. More generally, Meacham argues against the possibility of what he calls "autonomous chances", i.e. chance values which are objective but nevertheless different from the chance values the best system of the world assigns to events. Meacham's mentioned two different accounts of autonomous chances, which steam from different motivations.

The first account is that of Callender and Cohen (2010) that want to establish the independence of the special sciences from physics and so suggest an account of laws and chances for the special sciences which can be autonomous from those of physics. The second account is Hoefer's one (which has a more general motivation).

Meacham's attack on both these accounts is based on what he calls "the conflict problem":

> Chances are generally taken to place constraints on rational belief. All else being equal, if you know the chance of some event is 1/2, then your credence in that event should be 1/2. But if we have multiple autonomous chance theories, it seems like these different chance theories could impose conflicting constraints on rational belief. Call this the Conflicts Problem. (Meacham forthcoming, p. 2)

It is easy to see that "the conflict problem" just is Hoefer's conclusion. Meacham agrees with me that one potential way of "solving" the conflict problem is to take advantage of the admissibility clause in the PP, but rules this possibility out as unsatisfactory. Meacham writes:

> it's worth mentioning why one tempting way of modifying the chance-credence principle in order to avoid problematic prescriptions – adding an admissibility clause... doesn't look promising. This kind of proposal faces all three of the challenges sketched above – precisely charactarizing the resulting principle (and the notion of admissibility it employs), showing that the resulting principle avoids conflicts, and addressing motivational questions regarding the principle and the chances it employs. (Meacham forthcoming, p.16).

In the paper he only discusses, however, the third challenge. His argument is the following. Adding an admissibility clause to the PP must be motivated in the following sense: it must be shown that adding the admissibility clause to the PP is required in order for the PP to perform its role as the principle that captures the conceptual role "chance" plays. As should be clear from the discussion in section 2, I am sympathetic to this demand.

As mentioned, Meacham thinks that the admissibility clause is motivated in this sense when it comes to Lewis' first formulation of the PP. According to Meacham, the role of the admissibility clause in Lewis' first formulation is to make the second formulation of the PP (which neatly captures – according to Meacham - the conceptual role of "chance" and does not contain an admissibility clause) identical to the first formulation.

However, Meacham argues, an inclusion of an admissibility clause in Lewis' second formulation of the PP is unmotivated since the second formulation neatly captures the conceptual role chance plays. To support this claim Meacham considers several potential worries one might have concerning Lewis' second formulation of the PP that might motivate adding an admissibility cause to it and rules them out. One of these potential worries is the following one:

> The other potential worry one might have is that the Principal Principle is too strong without an admissibility clause... In particular,

> one might think that in 'crystal balls' cases, where (say) an agent get evidence about the outcomes of future events, her credence should not line up with the chances... (Meacham forthcoming, p.17)

Meacham's answer to this worry has a few steps, but for our discussion it is enough to concentrate on one of them (which is the one I reject). Suppose there are "infallible" crystal balls, argues Meacham, and let us imagine a rational agent that knows the best system of the world and also knows that a given crystal ball says that some future event, E, to which the best system of the world assigns some non-trivial chance, x, will occur. Still the agent needs, argues Meacham, evidence that the given crystal ball is indeed infallible. Now, since the best system of the world tells the agent that the chance of E is x, the agent clearly has no evidence that the crystal ball is infallible and so he should ignore its predictions and set his credence in E to be equal to x.

Here is Meacham's again:

> Suppose the crystal ball infallibly indicates that A will occur... Then either the agent's total evidence TK entails A, or it doesn't. If TK doesn't entail A, then the agent shouldn't heed the crystal ball's predictions are correct, since her total evidence doesn't give her reason to think the crystal ball's predictions are correct. So she should line up her credences with the chances, just as the Principal Principle says.

There is, however, a gap in the argument: How does it follow from "TK doesn't entail A" that the agent's total evidence "doesn't give her reason to think the crystal ball's predictions are correct"? Suppose, for example, a given crystal ball has given only accurate predictions in the past, including in cases in which it predicted an occurrence of an event to which the (known) best system of the world assigned a very low chance. In such a case the agent's total evidence does not entail the infallibility of the

crystal ball, but the agent does seem to have a good reason to think the crystal ball's predictions are correct.

Meacham's thought, I take it, was that to the extant the agent does have such good reasons they must be reflected in the chances the best system of the world assigns to the events predicted by the crystal balls. However, as we saw in the previous section, this is not necessarily true as the best system of the world cannot be the fittest system regarding all types of events. There will always be some type of events that exhibit some regularity, which is not reflected in the chances according to the best system. It might be the case that our agent will notice this regularity and if he does, it will be irrational of him to ignore it when setting his degrees of beliefs. The examples in section 3 shows that such cases are in fact not very rare in our world.

Thus, contrary to Meacham's position, there seem to be a clear motivation for including an admissibility clause in Lewis' second formulation of the PP: without such a clause, the second formulation is false. The sense in which it is false is that it does not accurately characterize the conceptual role "chance" plays. It is just not true that our concept of chance demands that after learning that the chance (now) of Bob opening the window is 0.001, learning that Bob will actually open the window is not evidence that Bob intended to open the window (a few seconds ago).

The discussion in the previous section shed further light on the motivation for including an admissibility clause in the PP. As was argued, the admissibility clause is needed in order to enable a rational agent overcome the limitations the simplicity consideration put on the best system of the world. While the best system of the world must give up on some of its fitness in order to gain in terms of simplicity, a rational agent's credence function is not bound under this demand. The admissibility clause is what allows a rational agent to overcome this limitation of the best system of the world. Now, this motivation for an inclusion of an admissibility clause holds under both formulations of the PP, not only under the first one.

Notice also that while Meacham took the second formulation of the PP to be the more fundamental one and argued that the role of the admissibility

clause is only to make the first formulation equivalent to the second, in his original paper, Lewis presented the conceptual relation between the two formulations the other way round. Lewis took the first formulation to truly capture "all that we know about chance" (Lewis 1980, p. 266) and introduced the second formulation (while being explicitly open to the idea that the two formulations are not equivalent due to the possible failure of the AHIT) only as a principle which is "easier to use" (Lewis 1980, p.277). Lewis also took the second formulation to enjoy "less direct intuitive support than the original formulation" (Lewis ibid).

I agree with Lewis that it is the first formulation that best captures the way we use the concept of chance both in our everyday uses and in scientific discourses. The second formulation is much less intuitive and the reason for that is that it is false: it assumes the AHIT which rules out some intuitive judgements we have regarding chancy events. The first formulation, however, needs an admissibility clause in order to be consistent. As explained in section 1, since this is so we must find a way to at least partly characterize the set of admissible propositions (to a given proposition). I have argued that the AHIT is not the right characterization to use, but I did not offer an alternative characterization (thus I have not provided answers to Meacham's first and second challenges). I intend to do that elsewhere.

Now, although chances are against me (judging by the current acceptance rate in philosophy journals), in case I will succeed in doing that, my intention will serve as a contrastive explanation for my offering such a characterization rather than not. Since this is so, my intention is (now) inadmissible to my actually doing so sometime in the future. Thus, my degree of belief that I will is pretty high. There is nothing irrational in that.

References

Callender, C, and Cohen, J. (2010), Special Sciences, Conspiracy, and the Better Best System Account of Lawhood, *Erkenntnis,* 73. pp. 427-447.

Elga, A. (2004), Infinitesimal Chances and the Laws of Nature, *Australasian Journal of Philosophy, 82, pp. 67-76.*

Hoefer, C. (2007). The Third way on Objective Probability: A Sceptic's Guide to Objective Chance, *Mind,* 116. pp. 549-596.

Hall, N. (1994), Correcting the Guide to Objective Chance, *Mind,* 103, pp. 505-517.

Hall, N. (2004), Two Mistakes about Credence and Chance, *Australiasian Journal of Philosophy,* 82 (1), pp. 93-111.

Lewis, D. (1980): A Subjectivist's Guide to Objective Chance, in R. C. Jeffrey (ed.), 1980, *Studies in Inductive Logic and Probabilities,* Vol. II, Berkeley: University of California Press, pp. 263-293.

Lewis, D. (1986), *Philosophical Papers: Volume II,* Oxford University Press.

Lewis, D. (1994), Humean Supervenience Debugged, *Mind,* 103. pp. 473-490.

Meacham, C. (2010), Two Mistakes Regarding the Principal Principle, *British Journal for the Philosophy of Science,* 61. pp. 407-431.

Meacham, C. (Forthcoming), Autonomous Chances and the Conflicts Problem, in *Asymmetries in Chance and Time,* edited by handfield and Wilson, Oxford University Press.

Percival, P. (2000), Lewis' Dilemma of Explanations under Indeterminism Exposed and Resolved, *Mind,* 109, pp. 39-64.

Smart, B. (2013), Is the Humean Defeated by Induction, *Philosophical Studies,* 162, 2, pp. 319-32.

Strevens, M. (1999), Objective Probability as a Guide to the World, *Philosophical Studies,* 95, pp. 243-275.

Opinion polling and election predictions

Abstract
Election prediction by means of opinion polling is a rare empirical success story for social science, but one not previously considered by philosophers. I examine the details of a prominent case, namely the 2012 US presidential election, and draw two lessons of more general interest:
1) *Methodology over metaphysics*. Traditional metaphysical criteria were not a useful guide to whether successful prediction would be possible; instead, the crucial thing was selecting an effective methodology.
2) *Which methodology?* Success required sophisticated use of case-specific evidence from opinion polling. The pursuit of explanations via general theory or causal mechanisms, by contrast, turned out to be precisely the wrong path – contrary to much recent philosophy of social science.

1. Introduction: metaphysics and methodology

Is systematic predictive success in social science possible? Many have given reasons why it is not, such as the fact that social systems are open systems, or that they exhibit reflexivity, or simply that there are too many variables needing to be modeled (Taylor 1971, Giddens 1976, Hacking 1995, Lawson 1997). In this paper I examine a notable case of predictive success so far relatively neglected by philosophers – namely election prediction by means of opinion polling – that seems to contradict these reasons.

Next, if successful prediction is possible, what makes that so? The lesson from the opinion polling case is that the most fruitful answer to this question is not metaphysical but rather is *methodological*. In particular, success or the lack of it was not predictable from the metaphysics of elections, which indeed in many respects remain unknown.[1] Rather, what was crucial was a certain methodological approach.

One popular methodological view, borne in part from pessimism about the possibility of prediction, has argued that the main aim of social science should instead be *explanation*.

1 Strevens (2005), for instance, gives metaphysical conditions for when explanation and prediction are possible in some complex systems. But it is unclear whether these conditions are satisfied in the elections case.

This latter can be achieved via the discovery of causal mechanisms, as urged by 'new mechanists' (Lawson 1997, Brante 2001), or else via the development of underlying theory (Elster 1989, Little 1991). Moreover, much of mainstream practice in economics – and other social science – is implicitly committed to this latter view: while all accept that rational choice models, for instance, might not be predictively successful, nevertheless they are held to provide 'understanding' or 'underlying explanation'.

A contrary view rejects this methodological emphasis on mechanisms or underlying theory (Cartwright 2007, Reiss 2008). One strand, motivated in part by detailed case studies of other empirical successes, has emphasized instead context-specific and extra-theoretical work. Theories and mechanisms play at most a heuristic role; empirical success requires going beyond them (Alexandrova 2008, Alexandrova and Northcott 2009).

The details of the opinion polling case turn out to endorse this second view. The reason is that, roughly speaking, while the extra-theoretical approach achieved prediction but not explanation, the theory-centred approach achieved neither. That is, a search for explanation not only yielded no predictive success, it also yielded no explanations. So the first view gives exactly the wrong advice.

I focus on the 2012 US presidential election, in which Barack Obama defeated Mitt Romney. I begin by describing the predictive success achieved by aggregators of opinion polls (section 2), before examining how this success was achieved (sections 3 and 4). In contrast, theoretical approaches to election prediction fared much worse (section 5). I then discuss their failure also at furnishing explanations (sections 6 and 7).

2. Predictive success and metaphysical criteria

The 2012 presidential campaign featured literally thousands of opinion polls. The most successful of all election predictors were some *aggregators* of these poll results. Famously, several successfully forecast the winner of all 50 states in the 2012 election, as well as also getting Obama and Romney's national vote shares correct to within a few

tenths of a percent.[2] This was a stunning success, arguably with few equals in social science. Nor was it easy – no one else replicated it, although many tried.[3] On the morning of the election the bookmakers had Romney's odds at 9/2, i.e. about 18%. Political futures markets such as InTrade had Romney's chances at about 28%. These market prices imply that common opinion was surprised by the outcome.[4]

Moreover, it is not reasonable to declare this success a mere fluke. First, the same poll aggregators have been successful in other elections too. And within any one election there have been many separate successful predictions, such as of individual Senate races or of margins of victory, which are at least partially independent of each other. Second, the aggregators' methods are independently plausible. It therefore behooves philosophers of social science to understand them.

Meanwhile, do elections satisfy the metaphysical criteria allegedly necessary for predictive success? It seems not. Presidential elections are clearly open systems in that they are not shut off from causal influences unmodelled by political science. They undoubtedly feature many variables. And they are clearly subject to reflexivity concerns: sometimes the mere publication of a poll itself influences an eventual election result; indeed there were several examples of this in the 2012 campaign. Yet despite such troubles, highly successful prediction proved possible nevertheless.

---

2 Four of the most successful aggregators were: http://votamatic.org, http://www.huffingtonpost.com/news/pollster/, http://election.princeton.edu/, and http://fivethirtyeight.blogs.nytimes.com/. The forecasting models for the first two of these were designed mainly by political science academics, the third by a neuroscience academic, and the last by a non-academic. Three of the four got every state right.

3 See http://delong.typepad.com/sdj/2012/11/war-on-nate-silver-final-after-action-report-the-flag-of-science-flies-uncontested-over-silvergrad-weblogging.html for a list of 47 examples of failure, with an emphasis on their suspicion of polling-based prediction.

4 There is an issue here about exactly what we are predicting and, thus, how we measure predictive success. After all, these market prices still had Obama as favorite, so why should we term them 'surprised' by his victory? In reply, first, besides the simple fact of the overall winner, there were also relevant additional facts: who won each state; and by how much did they win them? Odds-makers were not impressive with respect to these more detailed targets. Indeed, barring unlikely background assumptions, the details of the state-level results are hard to reconcile with a 28% chance of overall Romney victory. Second, the best poll aggregators' predictions were probabilistic, which makes it quite an intricate matter assessing who actually did best. (For analysis, see http://rationality.org/2012/11/09/was-nate-silver-the-most-accurate-2012-election-pundit/.) But there is no serious dispute that the odds-makers and many other predictors were not accurate.

3. The science of opinion polling[5]

In any opinion poll, the voting intentions of a sample serve as a proxy for those of a population. How might things go wrong, such that the sample will not be representative? The most well-known way, a staple of newscasts and introductory statistics courses alike, is *sampling error*: small samples can lead to misleading flukes. But sampling error is not the only source of inaccurate predictions and indeed is far from the most important one. Awareness of this crucial point lies at the heart of any serious election prediction.

To begin, a major issue for pollsters is to ensure that their samples are appropriately balanced with respect to various demographic factors. Suppose, for example, that two-thirds of interviewees were women. Since there was good reason to think that women were disproportionately likely to vote for Obama, it follows that such a woman-heavy sample would give misleadingly pro-Obama predictions. Polling companies would therefore *rebalance* such a sample, in effect putting greater weight on men's responses. Notice several things about such a rebalancing procedure.

First, it is quite different from sampling error.[6] In particular, if our sampling procedure over-selects women, then the error will not be reduced just by making the sample larger.

Second, sample rebalancing is clearly unavoidable if we wish to predict accurately. For this reason, every polling company performs some version of it.

Third, a poll's headline figures are therefore heavily *constructed*. They are certainly not the raw survey results. Exactly what and how much rebalancing is required depends on assumptions about the actual turnout on election day. For instance, in recent American

5 Although election prediction is the focus of this paper, opinion polls of course have many other uses too.

6 Lying in the background here are reference class issues. If we partition the population fine-grainedly enough, presumably even instances of sampling error will not be 'random'. But given the unavoidable cognitive and epistemic constraints facing polling scientists, choice of reference class is not arbitrary. And in practice the distinction between sampling and systematic error is of enormous importance to election prediction.

presidential elections typically there have been slightly more women than men voters, so it would be a mistake to rebalance the sample to exactly 50-50. The correct figure may not be obvious, it needing to be inferred from imperfect polling data about past elections, and moreover with some assessment about how patterns of turnout in the upcoming election may be different from those in previous ones. Accordingly, different polling companies may quite reasonably choose slightly different rebalancing procedures. The result is the phenomenon of 'house effects', when any particular company's polls may systematically favor one or other candidate compared to the industry average. When assessing the significance of a poll for election prediction, it is vital to be aware of this.

Fourth, the rebalancing issue is pressing because it applies to many other factors besides gender, such as: age; income; race; likeliness to vote; education; ownership of cellphones but not landlines; and home access to internet. Not only is the precise rebalancing procedure for each of these factors arguable, it is also arguable exactly which factors should be rebalanced for in the first place (see below).

In addition to random sampling error and systematic sampling bias, there are several other potential sources of error as well. There is space only to mention a couple here. One is the phenomenon of herding: at the end of a campaign pollsters – it is widely suspected – 'herd', i.e. report headline figures closer to the industry mean, presumably to avoid the risk of standing out as having missed the final result by an unusually large margin. Some sensitivity to this turns out to be optimal for accurate election prediction.[7] A second worry is simply that voters may change their minds between a poll and election day. This is the main reason why polls taken, say, six months before an election have a much poorer predictive record than do those taken closer to the time. Election predictions must therefore take into account a poll's date too.

4. Poll aggregation

7 For evidence of herding's significance, and references, see: http://fivethirtyeight.blogs.nytimes.com/2012/10/11/oct-10-is-romney-leading-right-now/, http://votamatic.org/pollsters-may-be-herding/, and http://themonkeycage.org/2012/09/26/robo-polls-a-thumb-on-the-scales/

Turn now to the aggregation of polls, which represents a second layer of method, quite distinct from that required for a single poll. Historically, poll aggregation has had a better predictive record than using individual polls alone. One obvious reason is that aggregation increases effective sample size and therefore reduces sampling error. A typical individual poll may have 95% confidence intervals of 3 or 4%; the confidence intervals for an aggregation of eight or ten polls, by contrast, are typically 0.75 or 1%.[8] But it is a different story for the other possible sources of error. Mere aggregation is no cure for those, because it might be that they bias all polls – and hence the aggregate of polls – in a similar way.[9]

What, then, does account for aggregation's superior predictive success? In part, it is indeed simply the reduction of sampling error. But it is not just that; it is also that *sophisticated* aggregation can mitigate the other sources of error too. This explains why the best aggregators beat simple averaging of the polls. It is instructive to consider a couple of methodological issues in more detail.

4.1) State versus national polls

One feature of the 2012 US presidential campaign was a divergence between state and national polls. By combining opinion polls for individual states, making due allowance for population and likely turnout, it is possible to calculate an implicit figure for the vote share across the country as a whole. When this was done, there was a surprising inconsistency: the state polls implied that Obama was ahead at the national level, but the national-level polls showed him behind. The divergence was at least three percentage points. What to do? Simple averaging was no answer, because the inconsistency was true of the polls' averages too.

One possible cause of the divergence was that it was just sampling error – confidence intervals are sufficiently wide that there is a non-negligible chance of this. However, a similar discrepancy had persisted for much of the campaign, rendering this explanation

8 There were rarely more than eight or ten polls of a single area in a single time period.
9 Although house effects will, by definition, tend to cancel out, still it might be that the best sample rebalancing procedure is an outlier relative to the industry average.

implausible. Another possibility was that Obama's votes were disproportionately concentrated in heavily-polled swing states. But this explanation turned out to be implausible too: first, it required disproportionately good Romney polling in non-swing states, but this had not occurred. Second, it seemed unlikely anyway given that, demographically speaking, swing voters in Ohio or Virginia are much the same as those in Texas or California, so why should their voting intentions be systematically different? – after all, both campaigns were spending similar amounts in the swing states. Third, such a pattern is uncommon historically.

There therefore seemed little prospect of reconciliation; instead, it boiled down to preferring one of the state or national polls to the other. In favor of the latter: national polls tend to have larger sample sizes and to be run by more reputable firms. But on balance, there were better reasons to prefer the state polls instead. First, there are many more of them, suggesting that sampling error is less likely. Second, some of the other sources of error are arguably less likely too. In particular, herding effects will likely occur relatively independently in different states. As a result, that source of error for state polls will likely cancel out at the level of national polling numbers. Third, historical evidence again: when the two have conflicted in previous elections, typically the state polls have proven better predictors than have national ones.

The take-away is to emphasize the value added by sophisticated poll aggregation. Simply averaging the polls was not enough. Neither was it optimal just to split the difference between state and national polls symmetrically. Instead, more sophisticated analysis was required.

4.2) Could all the polls have been wrong?

By the end of the 2012 campaign, it was clear that if the polls were right then Obama would win. Romney's only hope by then was that the polls were systematically skewed against him. Thus, all turned on whether the polls could indeed be so skewed. Once again, simple averaging is no help here, since the issue at hand is not *what* the polls said but rather whether we should *believe* what they said.

The historical record suggested it was unlikely the polls were skewed enough to save Romney.[10] Confidence in this conservative verdict was strengthened by the absence in 2012 of factors that have marked systematic polling errors in the past, such as a high number of voters declaring themselves 'undecided', or a significant third-party candidate. Given that it was the end of the campaign, there was also little reason to expect a large change of voters' minds before election day – especially given also the record levels of early voting. Finally, the number of polls involved and the size of Obama's lead made sampling error too an implausible savior for Romney.

The only remaining source of error was therefore sample rebalancing. In particular, was there some procedural skew, common across many or all polls, which had been mistakenly depressing Romney's figures? There was little evidence of a significant 'Bradley effect', i.e. where polls overrated Obama because respondents were reluctant to state their opposition to him for fear of seeming racist.[11] But a different possibility was much discussed. It concerned whether polling companies should rebalance samples according to party affiliation. American voters self-identify as one of Democrat, Republican or Independent. If a polling sample were, say, disproportionately composed of Democrats, that would yield a skewed pro-Obama result. In 2012, that was exactly the accusation: polls showed that Romney had a big lead among Independents, and critics charged that Obama came out ahead overall only because the polls were 'over-sampling' Democrats. That is, the proportion of Democrats in samples was charged to be disproportionately high in light of exit polls from previous elections and other considerations.[12]

The key methodological issue is whether party affiliation is a stable population variable that should be adjusted for in the same manner as age or gender, or whether instead it is an unstable variable that is merely an attitude and often just an *effect* of voting

10 http://fivethirtyeight.blogs.nytimes.com/2012/11/04/nov-3-romneys-reason-to-play-for-pennsylvania/
11 http://www.fivethirtyeight.com/2008/08/persistent-myth-of-bradley-effect.html
12 http://www.redstate.com/2012/10/26/why-i-think-obama-is-toast/

preferences. If the latter, then the party of whomever is the more popular candidate may be 'over-sampled' simply because a voter's party self-identification is influenced by their voting intention in a way that their age or gender cannot be. If so, then it is distorting to rebalance for stated party affiliation; but if not so, then it is distorting *not* to. Standard industry practice had been the former, i.e. not to rebalance for stated party affiliation. Predicting correctly who would win the presidency turned critically on whether this practice was correct. Was it?

Again, simply averaging the polls was no help. One piece of evidence gives a flavor of the more detailed kind of analysis required. Across different polls of a particular state, with similar headline figures, there was typically a strong positive correlation between Romney's lead among Independents and the proportion of voters self-identifying as Democrats.[13] The inference from this is that party self-identification is an unstable variable. For various reasons, a given Obama voter might self-identify as Independent in one poll but as Democrat in a second. As a result, in the first poll there are fewer Democrats and Romney's lead among Independents is lower, whereas in the second both are higher – hence the positive correlation between the two. The important thing from a prediction point of view is whether Obama is leading overall in both polls – as, in swing states, he indeed was.

5. Failure of the theory-centred approach

The details of poll aggregation show clearly the case-specific nature of its methods. The alternative is to focus instead on 'fundamentals', i.e. on variables that might shed light on election results generally not just case-specifically, such as economic conditions, the perceived extremism of candidates, incumbency, and so forth. There is a literature in political science on election prediction that aims to furnish just such generalizable models.[14] How does it fare?

13 http://www.huffingtonpost.com/nick-gourevitch/romney-lead-with-independents_b_2058290.html

14 Influential contributions include Fair (1978), Campbell and Wink (1990), Hibbs (2000), Abramowitz (2008), and Lewis-Beck and Tien (2008). Montgomery et al (2012) averages these and other models to achieve the best forecasting success of all.

Conveniently, it too has focused on US presidential elections. The sample size is relatively small, as fewer than 20 elections have good enough data. This creates a danger of overfitting. In response, models typically feature only a small number of variables, most commonly economic ones such as growth in GDP, jobs or real incomes.[15] Sensibly, they are estimated on the basis of one part of the sample and then tested by tracking their predictive performance with respect to the rest of the sample.[16] Even then, there remains a risk of overfitting – if a model predicted the first few out-of-sample elections quite well, will its success continue in future elections? Moreover, even if a model does successfully predict past elections, there is no guarantee the political environment is so stable that the model will remain valid in future too.

These caveats noted, it is true that the models do have a little success. On one estimate, the best ones' average error when predicting the incumbent party's share of the vote is between 2 and 3%.[17] But this is not quite as impressive as it might initially sound: first, for our purposes it is something of a cheat, in that one of the variables in by far the highest weighted model – Abramowitz 2008 – is a polling result, namely presidential approval rating. So the success is not achieved purely by fundamentals. Second, a 2-3% average error corresponds to an average error when estimating the *gap* between the leading two candidates of about 5%. And third, vote shares rarely deviate all that much from 50% anyway, so they are quite an easy target – indeed, another estimate is that economic variables account for only about 30-40% of the *variance* in incumbent party vote share.[18] Overall, the models do not predict individual election results very reliably.

15 Literally thousands of economic variables could plausibly be deemed relevant, not to mention many non-economic ones too.

16 Of course, there is a long history of debate within philosophy of science about the relative epistemic merits of novel prediction versus retrospective accommodation of the facts. Defenses of the latter typically emphasize how theory may have been developed or tested independently of the particular accommodation, how background knowledge may leave the main epistemic task mere calibration of an agreed functional form, or how a lack of any plausible alternative explanations tells in favor of the one that we do find. But none of these defenses apply well to the election prediction case, justifying the literature's concentration here on prediction.

17 See http://www.brendan-nyhan.com/blog/2011/11/a-comparison-of-presidential-forecasting-models.html for discussion and references.

18 http://fivethirtyeight.blogs.nytimes.com/2011/11/16/a-radical-centrist-view-on-election-forecasting/

On many occasions, they even get wrong the crude fact of which candidate won. For accurate prediction, it is necessary to incorporate the results of opinion polls.

6. Explanation

Why did Obama win? Answering this requires identification of an event's causes.[19] That in turn requires a verified theory or causal model. The problem is that nobody – from either approach – has managed to produce one.

On the polling side, in a trivial sense Obama's victory is 'explained' by the fact that, as revealed by aggregators, on the eve of the election a majority of the electorate were minded to vote for him. But, of course, for most investigative purposes a deeper explanation is required, in particular one that might apply to other elections too. Poll aggregation provides none.[20]

On the fundamentals side, if their models had fared better they would have provided exactly the explanations that polling aggregation does not. After all, that is precisely the motivation for theory-centred methodology. Thus, say, we might have been able to explain that Obama won because of positive GDP and jobs statistics in the preceding two quarters. Unfortunately, though, the fundamentals models are not predictively accurate.

Can they nevertheless provide us with explanations anyway? The argument would be that they have truly identified relevant causes. It might be postulated, for instance, that GDP or stock market growth does causally impact on voter preferences and thus on election outcomes. True, other causes impact too and so the models do not explain the outcomes fully nor predict them accurately, but that still leaves room for the claim that they explain them 'partially' by correctly identifying *some* of the causes present.[21]

19 Following the literatures under consideration here, I focus on *causal* explanation. I do not mean to rule out the possibility of other forms of explanation.

20 It is true that some proximate explanations of the election outcome can be *tested* by careful observation of movements in the polls. For instance, the impact of Obama's weak performance in the first presidential debate, or of Hurricane Sandy in the campaign's final week, can be assessed in this way. But testing an explanation is not the same as providing one.

21 See Northcott (2013) for more on the relevant sense of partial explanation.

But, alas, even this claim is dubious. First, the different models cite different variables. Abramowitz's, for instance, cites GDP growth, presidential approval rating, and a complex treatment of incumbency; Hibbs's though cites growth in real disposable income and the number of military fatalities abroad. Even among economic variables alone, some models cite GDP, some household incomes, some jobs data, some stock market performance, and so on. There are many different ways to achieve roughly the same limited predictive success, which shakes our faith that any one way has isolated the true causal drivers of election results. Perhaps the small sample size relative to the number of plausible variables makes this problem insoluble.

The second reason for pessimism is that, elsewhere in science, a standard response to predictive failure is to test putative causes in isolation. As it were, at least we achieve predictive success in the isolated test. But unfortunately such experiments are impossible in the case of election predictions. So as well as predictive failure at the level of elections as a whole, the causal factors picked out by the models have not earned their empirical keep by other means either.

The upshot is that we have no warrant for asserting that we have found even some of the causes of election outcomes, and therefore no warrant for claiming even partial explanations. Thus the basic conclusion stands: we have not achieved any explanation of election outcomes, and so the original motivation for turning to fundamentals models is frustrated.

7. Transportability

Are the predictive successes of one election transferrable to another? That is, will a similar polling aggregation strategy work elsewhere? For US presidential elections, it seems that the answer is 'yes' – witness the success of many of the same polling aggregators in 2008. However, it is a different story for other elections, such as US congressional elections or elections in other countries.[22]

22 The http://fivethirtyeight.blogs.nytimes.com/ predictions of these two in 2010, for instance, were notably less successful.

The reason is precisely the case-specific nature of polling aggregation – for the best aggregation does not rely only on polls. It must also factor in features such as whether an election is a single national vote or split into many smaller constituencies, whether there are two or many major political parties, whether the voting system is first-past-the-post or proportional representation, one-shot or multi-round, and so on. The implications of a poll for election prediction depend on just such factors. It has also proved profitable to moderate polling results by considering what result should be 'expected', given various local demographic and historical factors. The details of just how to do this are important – and inevitably highly case-specific.

Perhaps even more significantly, the earlier nuances, namely adjudicating state versus national polls and whether polls might be systematically skewed, could also only be resolved by case-specific knowledge. There are many similar examples, such as the extent of regression to the mean to be expected if one candidate is 'surprisingly' far ahead at an early stage of the campaign, or after party conventions or presidential debates. Such knowledge is crucial, but typically it is transferrable to new elections only imperfectly if at all.

So a serious polling aggregator must build a new election prediction model each time. This lack of transportability is really just the flipside of two facts familiar from above: first, that no one has achieved satisfactory causal explanations – not even the poll aggregators. And second, that predictive success requires case-specific knowledge rather than a search for generalizable causal mechanisms or theoretical underpinnings.

8. Conclusion

How can we make progress, i.e. predict election results even better? It is clear that improving the models of fundamentals is an unpromising route. Rather, progress will be made in the same way as it has been made in the last few years – by doing polling aggregation better. This might involve getting better polling data, analyzing that data better, or understanding better how the implications of that data depend on local

peculiarities – in other words, by developing the case-specific, extra-theoretical components of prediction for each application anew. Although transportable explanations are elusive, predictive success need not be; what is clear, though, is that misplaced context-free theory offers neither. It seems there are no short-cuts in social science.

References

Abramowitz, A. (2008). 'It's About Time: Forecasting the 2008 Presidential Election with the Time-for-Change Model', *International Journal of Forecasting* 24: 209-217.

Alexandrova, A. (2008). 'Making models count', *Philosophy of Science* 75, 383-404.

Alexandrova, A. and R. Northcott (2009). 'Progress in economics', in D. Ross and H. Kincaid (eds) *Oxford Handbook of Philosophy of Economics*, Oxford, 306-337.

Brante, T. (2001). 'Consequences of Realism for Sociological Theory-Building', *Journal for the Theory of Social Behaviour* 31, 167–94.

Campbell, J. and K. Wink (1990). 'Trial-Heat Forecasts of the Presidential Vote', *American Politics Quarterly* 18 (3): 251–69.

Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge.

Elster, J. (1989). *Nuts and Bolts for the Social Sciences*. Cambridge.

Fair, R. (1978). 'The Effect of Economic Events on Votes for President', *Review of Economics and Statistics* 60, 159-173.

Giddens, A. (1976). *New Rules of Sociological Method: a Positive Critique of interpretative Sociologies*. London: Hutchinson.

Hacking, I. (1995). 'The Looping Effect of Human Kinds', in D. Sperber, D. Premack and A Premack (eds) *Causal Cognition an Interdisciplinary Approach*, Oxford.

Hibbs, D. (2000). 'Bread and Peace Voting in US Presidential Elections', *Public Choice* 104, 149–180.

Lawson, T. (1997). *Economics and Reality*, London: Routledge.

Lewis-Beck, M. and C. Tien (2008). 'The Job of President and the Jobs Model Forecast: Obama for '08?', *PS: Political Science and Politics* 41, 687-90.

Little, D. (1991). *Varieties of Social Explanation*, Boulder, CO: Westview.

Montgomery, J., F. Hollenbach and M. Ward (2012). 'Ensemble Predictions of the 2012 US Presidential Election', *PS: Political Science and Politics* 45, 651-654.

Northcott, R. (2013). 'Degree of explanation', *Synthese* 190, 3087-3105.

Reiss, J. (2008). *Error in Economics: Towards a More Evidence-Based Methodology*. Routledge.

Strevens, M. (2005). 'How are the sciences of complex systems possible?', *Philosophy of Science* 72, 531-556.

Taylor, C. (1971). 'Interpretation and the Sciences of Man', *Review of Metaphysics* 25, 3-51.

All cited URLs were accessed in August 2013.

July 18, 20, 2014

Curie's Truism

John D. Norton[1]
Department of History and Philosophy of Science
Center for Philosophy of Science
University of Pittsburgh
Pittsburgh PA USA 15620
http://www.pitt.edu/~jdnorton

Curie's principle asserts that every symmetry of a cause manifests as a symmetry of the effect. It can be formulated as a tautology that is vacuous until it is instantiated. However instantiation requires us to know the correct way to map causal terminology onto the terms of a science. Causal metaphysics has failed to provide a unique, correct way to carry out the mapping. Thus successful or unsuccessful instantiation merely reflects our freedom of choice in the mapping.

1. Introduction

When Pierre Curie (1896) introduced the principle that now carries his name, his concern was a quite specific problem in crystallography. The properties of a crystalline substance supervene on the atomic structure of its crystalline lattice. Hence those properties must respect the symmetries of the lattice. If, in addition, the lattice is subject to external influences such as an electric or magnetic field, the symmetries to be respected reduce to those common to the lattice and external influence. This last remark is the substance of Curie's observation.

Curie expressed it as one of a number of "propositions" in the general language of cause and effect.[2]

[1] I thank my co-symposiasts, Elena Castellani, Jenann Ismael and Bryan Roberts for stimulating discussion.

> When certain causes produce certain effects, the symmetry elements of the causes must be found in the their effects.

This proposition continues as a basic supposition of crystallography. The generality of its form, however, has led it to appear in other sciences, such as structural geology. (See Nakamura and Nagahama, 2000). It has also entered the philosophy of science literature.

If one seeks an ever-elusive principle of substantial content in the metaphysics of causation, one might be tempted to identify this principle. As Brading and Castellani (2013) point out, it does appear to be a straightforward application of Leibniz's principle of sufficient reason. A symmetry expresses an indifference in a cause. We should expect that same indifference in the effect, since we lack a sufficient reason for it to be otherwise.

Appealing as this vindication of causal metaphysics may seem, the principle's status in the literature is fraught. A straightforward macroscopic account of spontaneous symmetry breaking is a *prima facie* counterexample. An isotropic ferromagnet, on cooling past its Curie (!) point, acquires a magnetization in some random direction. (The example is much disputed. See Ismael, 2007, §7; Castellani, 2003; and Earman, 2004.) Chalmers (1970, p. 134) allows that Curie's principle may be irrefutable, since we might overturn any counterexample by seeking some as yet undiscovered asymmetry. He also reports (p.133) Freundenthal's suspicion that the generality of the principle depends on its being "necessarily vague." The supposition that asymmetry can only come from asymmetry is falsified, van Fraassen (1989, p. 240) asserts, by indeterministic processes. Ismael (§§ 2, 3 and 6) responds that the principle is a demonstrable truth for both deterministic and indeterministic systems. Belot (2003, pp. 404-405) and, more forcefully, Roberts (2013) have described counterexamples to Curie's principle. In short, there is no consensus on the status of Curie's principle. It is all of an irrefutable, metaphysical necessity with counterexamples; a demonstrable truth; an empirical falsehood; and an overreaching vagueness.

My purpose in this paper is to identify precisely why Curie's principle engenders such a proliferation of opinions. I will argue that Curie's principle is a demonstrable truth, but merely as an easy tautology. Its success or failure in science depends entirely on whether it is instantiated in some system. Whether it is instantiated depends in turn on how we interpret elusive terms like

---

[2] Curie (1894, p. 127); translation Brading and Castellani (2003. p. 312).

"cause" and "effect." There is sufficient pliability in our interpretation of causal language to make the principle a truism, when it does turn out to be true. That is, it is a self-evident truth, but one whose truth is attained cheaply through a pliability in the meaning of causal terminology.

The pliability in our interpretation of causal language arises from the failure of causal metaphysics to deliver unequivocal meanings. Elsewhere (Norton, 2003, 2007, 2009), I have argued that the metaphysics of causation supports no independent, empirical principles of universal scope. Rather successful causal talk in science is merely the opportune attachment of causal labels to terms in the propositions of a science, without in any way restricting their content. These same concerns apply here. Depending on how we construe the notions of cause and effect, we can render Curie's principle a truth of a selected application in science or not.

Each of the proliferating opinions of Curie's principle arises by emphasizing one or other aspect of these success and failures of instantiation. A sense of the pervasive truth of the principle comes from the fact that familiar construals of cause and effect enable successful instantiation. In some cases, other construals are so contrived as that we see no alternative. This is a purely fortuitous alignment of our causal prejudices with the case at hand. We mistake that accident as a manifestation of a deep truth of universal scope.

A sense that the principle is banal and its truth cheaply won, I will suggest, derives from an implicit recognition that these construals are not necessities. No higher principle precludes us using different ones that may lead the principle to fail. Finally, a sense that the principle is a falsehood stems from a recognition of the natural construals of cause and effect that preclude its instantiation.

In Section 2, I will give a more precise statement of the principle as tautology, a demonstrable truth. In Section 3, success or failure of the principle will be characterized as success or failure to instantiate the tautology. The remaining Sections 4 and 5 will provide illustrations of the failure of the principle to be instantiated in a context in which it is generally assumed to succeed (deterministic theories); and illustrations of the success of the principle in contexts in which it is normally supposed to fail (indeterministic theories).

2. Curie's Principle Formulated as a Lemma

Informally stated, Curie's principle requires that any symmetry of a cause manifests as a symmetry of the effect. To convert this into a demonstrable proposition, we need to make the notions invoked a little more precise.

Causes and their symmetries. The set of possible causes $\{C_1, C_2, C_3, \ldots\}$ admits a group G_C of symmetry transformations $\{S^C_1, S^C_2, S^C_3, \ldots\}$ such that any symmetry S^C_i acting on any cause C_k satisfies $C_k = S^C_i \cdot C_k$.

Effects and their symmetries. The set of possible effects $\{E_1, E_2, E_3, \ldots\}$ admits a group G_E of symmetry transformations $\{S^E_1, S^E_2, S^E_3, \ldots\}$ such that any symmetry S^E_i acting on any effect E_k satisfies $E_k = S^E_i \cdot E_k$.

This characterization is sparse. More would be needed if any of the details of the symmetry transformations are to be displayed.[3] However it is not required here since these details will prove irrelevant to what follows.

The causes C and effects E carry these names since they are related by functional determination. For Curie, the cause was a supervenience base of the crystal lattice and imposed fields; the effect was a property fixed by it synchronically. In the philosophy of science literature, the cause is most commonly an initial state and the effect is the state to which it evolves under some rule of time evolution. The general relation is:

Causes determine effects. There is a functional relation of dependence of effects on causes. That is, there is a function f, such that for each cause C_i there is a unique effect $E_i = f(C_i)$.

The dependence function must obey one restriction if it is to figure in a statement of Curie's principle: it must preserve any symmetry present in the cause when it maps causes to effects. This will be formulated more precisely as (CP2) below.

---

[3] For example, a cause might be the distribution of certain properties over a base space. The symmetry would map points in the base space to other points and carry the properties along in such a way that the final distribution is the same as the initial.

Curie's principle can now be formulated as a simple lemma, that is, a simple "if…then…" proposition:

Curie's Lemma[4]		
IF	(CP1) *Symmetry of cause*.	Causes C_i admit symmetries G_C.
	(CP2) *Determination respects symmetries*.	If causes C_i admit symmetries G_C and are mapped to effects $E_i = f(C_i)$ then there exists a symmetry group G_E that is isomorphic[5] to G_C.
THEN	(CP3) *Symmetry of effect*.	The effects $E_i = f(C_i)$ admit symmetries G_E isomorphic to G_C.

[4] This formulation fails to capture the informal intuition that the symmetry of the effect should be *produced* by the symmetry of the cause. It does not preclude the case of a cause that admits a symmetry SO_3 in space, while the effect admits no such symmetry in space but, coincidentally, an SO_3 symmetry in an internal space, not present in the cause. The formulation here of the tautology is good enough for the analysis that follows since that same analysis would apply to any augmented tautology.

[5] That is, there is a bijection between G_C and G_E that preserves group operations. Demonstrations of Curie's principle may assume that the same symmetry transformation S can act on both causes and effects without specifying the sense of sameness (e.g. Ismael, 1997, p. 169) The formulation of (CP2) here is more complicated to avoid this difficulty.

3. The Meaning of Success and Failure

It is clearly heavy-handed to lay out the principle in the form of the last section. For there is little substance to it. It is a tautology implementing as an easy modus ponens "A; if A then B; therefore B." That simplicity does make precise the sense that the principle somehow has to be true. For whenever a cause admits a symmetry and the rule of causal determination respects that symmetry in the precise senses of (CP1) and (CP2), then the symmetry must reappear in the effect as a matter of elementary logic.

Once we have formulated Curie's principle as a tautology, then its truth is automatic. How can there be any question of it succeeding or failing? That success or failure depends on whether the tautology is instantiated by some system of interest; that is, whether that system of interest provides a model of the tautology in the usual semantic sense in logic. Successes or failures of Curie's principle then depend entirely on how we map the terms appearing in its statement to system of interest.

Successes of Curie's principle arise when we perform the mapping so that (CP1), and (CP2) are verified. Failures of Curie's principle arise when we perform the mapping so that they are not verified. These facts powerfully restrict our analytic options. Any success of the principle must be traced back to this mapping verifying (CP1) and (CP2). Any failure of Curie's principle must be traced back to this mapping failing to do so.

We now see why the principle is a truism, when the instantiation succeeds. That just means that it is a pliable truth whose successful application to some system comes cheaply. It arises directly from the pliability of our mapping of the terms cause, effect and causal determination into the terms of the specific case at hand.

There will be many ways to carry out the mapping. When Curie's principle is applied to cases of deterministic time development, the natural mappings typically yield success. When indeterministic time developments are considered, however, the natural mappings do not. In particular, indeterministic time evolutions give rules of dependence that tend to violate symmetries. Hence successes of Curie's principle are normally associated with deterministic time development and failures with indeterministic time development.

The main claim of this paper, however, is that this association is happenstance. There is no higher principle that dictates which mapping is correct. What decides the mapping used is familiarity, comfort and, ultimately, our whim. The sections that follow will illustrate different

mappings that bring an unexpected failure of Curie's principle for a deterministic system and unexpected successes of Curie's principle for indeterministic systems.

3. Failure in a Deterministic Theory

Determinism alone cannot be sufficient to ensure Curie's principle. Some extra condition like (CP2) is required. This is shown by a toy example: highly symmetric causes $C_1, C_2, \ldots$ are mapped one-one to effects $E_1, E_2, \ldots$, each of which has no symmetry whatever. Curie's principle fails, since (CP2) is not verified. Below is a more realistic failure and a contrasting success.

3.1 Galileo's Law of Fall (failure)

A body with initial horizontal (x direction) velocity **v**(0) falls vertically (-z direction) with constant acceleration g. It is mapped as follows:

cause	The body at the instant t=0 moving with horizontal velocity **v**(0).
effect	The parabolic trajectory in the x-z plane; a compounded horizontal and vertical motion.
rule of dependency	Galileo's law of fall, expressed as $d\mathbf{v}(t)/dt = -g\mathbf{k}$, where **k** is a unit vector in the z direction.

Curie's principle fails for Galileo's law of fall, when the causal notions are mapped as indicated. The symmetries of the cause are all spatial rotations and mirror reflections that preserve **v**(0). However, as shown in Figure 1, the effect does not manifest these symmetries. Spatial rotations about **v**(0) are not symmetries of the parabolic trajectory of the effect.

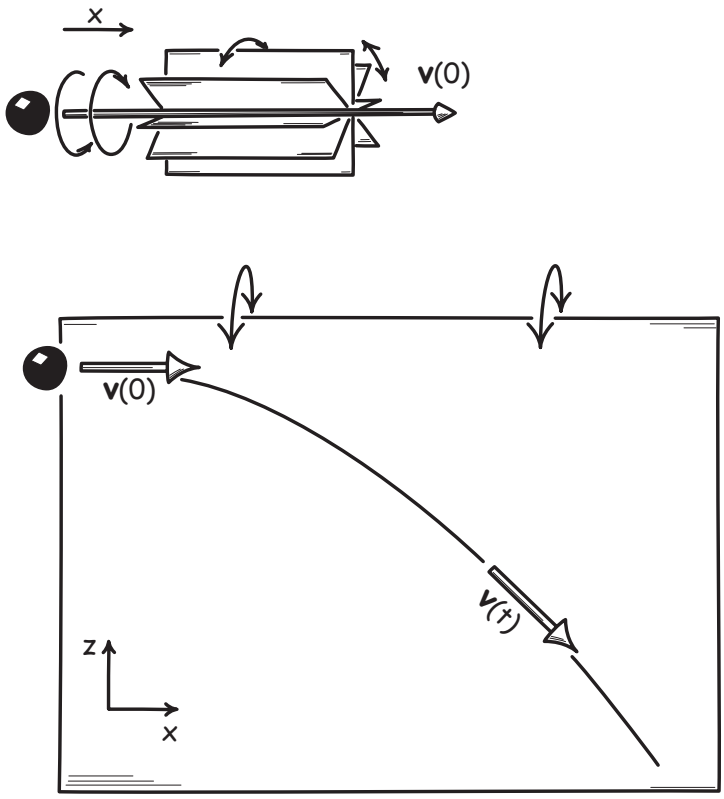


Figure 1. Symmetries of Galileo's Law of Fall

The reason for the failure is that condition (CP2) of the lemma is not verified. Galileo's law of fall does not preserve the full symmetries of the initial state. It introduces a vertical motion that violates the rotational and most mirror symmetries of the initial state about the axis of **v**(0).

3.2 Fall in a Gravitational Field (success)

There is an easy way to restore Curie's principle to the law of fall. We say that Galileo's law of fall, as expressed in Section 3.1, does not fully represent all the relevant causal processes. It introduces a preferred direction of space, the z direction, which is distinguished as vertical. We should, the restoration says, give the physical reason for this distinction. We now know that it is the presence of a gravitational field: $\varphi = gz$. Galileo's law of fall should be replaced by the Newtonian field version:

$$d\mathbf{v}(t)/dt = -\nabla\varphi = -g\mathbf{k},$$

We now map the augmented example as:

cause	The body at the instant t=0 moving with horizontal velocity $\mathbf{v}(0)$; and the gravitational field $\varphi = gz$.
effect	The parabolic trajectory in the x-z plane; a compounded horizontal and vertical motion.
rule of dependency	Galileo's law of fall, expressed as $d\mathbf{v}(t)/dt = -\nabla\varphi = -g\mathbf{k}$, where $\mathbf{k}$ is a unit vector in the z direction.

With the augmentation of the gravitational field and these new mappings, Curie's principle succeeds. The symmetries of the cause are now greatly reduced. They are merely the symmetries common to the initial velocity $\mathbf{v}(0)$ and the gravitational field $\varphi = gz$. That is just the single mirror reflection that preserves the x-z plane in which the initial velocity $\mathbf{v}(0)$ is found. This reduced symmetry is respected by Galileo's law of fall. The same symmetry now manifests in the effect, for the parabolic trajectory of fall is fully contained within this x-z plane.

3.3 Which is the Real Cause?

One might be tempted to dismiss the failure of Curie's principle in the first case as arising from an imperfect identification of the causes. Correctly identify the *real* cause as including the asymmetric field of the second case and then Curie's principle succeeds.

The temptation should be resisted. It rests on the presumption that asymmetries *have* to be included in causes and cannot be included in the rule of dependency.[6] That presumption conflates a familiarity with a necessity. Asymmetries in rules of dependencies do often later prove to result from other processes still; and that discovery may enable the asymmetry to be moved from the rule to the cause; and it may be done in a way that conforms with (CP1) and (CP2). However there is no necessity that it must always be so. A deterministic rule of dependency that breaks symmetries is unusual among physical laws, not impossible. A law of fall that includes a preferred direction in space contains no incoherence. Indeed a more elevated

[6] Ismael (1997, p. 171) seems to defend this view.

example has been a fixture in the standard model of particle physics for half a century. The weak interaction violates spatial parity conservation. To formulate the physical laws governing the weak interaction, we must introduce a preferred handedness into space, much as formulation of Galileo's law of fall requires identification of a preferred direction in space.

4. Success in Indeterministic Theories

The most mentioned failures of Curie's principle involve indeterministic time evolutions. Two examples are presented here. Depending on the mappings used, we can render them as successes or failures of Curie's principle.

4.1 A Probabilistically Stochastic Theory: Radioactive Decay

Consider the radioactive decay of an atom. To be specific, take the alpha decay of a heavy atom. The decay product, the alpha particle, will be projected isotropically in space; or at least it will be if we follow a Gamow-style model of alpha decay as the quantum tunneling of a particle from a spherically symmetric potential well. It will be governed by the law of radioactive decay, which asserts that the probability of decay in some small time interval dt is λdt, where λ is the decay constant. It follows that the probability density of the alpha particle being projected in angular direction θ, ϕ at time t is given by[7]

$$\rho(\theta, \phi, t) = (4\pi \sin\phi / \lambda) \exp(-\lambda t)$$

This density distributes the probability of projection isotropically, that is, uniformly over the angular directions θ, ϕ. The symmetry is then broken by realization of one direction of projection.

Here are two mappings of the causal notions. One leads to failure of Curie's principle; one leads to success of Curie's principle.

[7] θ, the longitude and ϕ, the co-latitude, are the standard angular coordinates of a spherical coordinate system.

	Curie's principle fails	*Curie's principle succeeds*
cause	Radioactive atom at t=0.	Radioactive atom at t=0.
effect	Specific radioactive decay event in which an alpha particle is ejected at time t in direction θ, ϕ.	Probability distribution $\rho(\theta, \phi, t)$ that an alpha particle is ejected at time t in direction θ, ϕ.
rule of dependency	Spatially isotropic law of radioactive decay; symmetry breaking projection.	Spatially isotropic law of radioactive decay.

The "fails" column has the normal mapping. The cause, the radioactive atom, is spherically symmetric in space. The effect, the emission of an alpha particle in a particular spatial direction, violates this symmetry. Curie's principle fails to be instantiated. The failure derives from the failure of the mappings to verify (CP2). For the effect, the particular time and direction of the decay, is not functionally dependent on the cause, the state of the atom at t=0; and the rule of dependency allows the decaying alpha particle to move in a particular direction, contrary to the symmetry of the atom.

The "succeeds" column indicates another mapping. The effect is not the individual decay event, but the probability distribution to which it conforms. With this mapping, we can quickly verify that Curie's principle succeeds. The spherical symmetry of the cause, the radioactive atom, is respected by the spatially isotropic law of radioactive decay. The new effect, the probability distribution $\rho(\theta, \phi, t)$, manifests the spherical symmetry of the cause. Replacing the individual decay event by the probability distribution averages away the spatial anisotropy of particular effects, allowing (CP2) to be verified.

4.2 A Non-Probabilistic Indeterministic Theory: The Dome

The sort of indeterminism manifested in radioactive decay is limited in the sense that the undetermined futures must conform to a probability distribution. There are many examples of a more extreme failure of determinism. The physics is indeterministic and, crucially, it provides no probability distributions to which the many admissible futures must conform. Yet we shall see that this more severe form of indeterminism is just as hospitable to Curie's principle.

To make the analysis concrete, consider the simplest example of this sort of indeterminism: the "dome" within ordinary Newtonian physics. A radially symmetry dome has radial coordinate r along the surface of the dome and angular coordinate θ. It sits in a uniform gravitational field and is shaped so that a point at r is depressed vertically below the apex by a distance $h = (2/3g)r^{3/2}$, where g is the acceleration due to gravity. At time t=0, a point mass that can slide frictionlessly over the dome surface is motionless at the apex, r=0. It is easy to show (Norton, 2003, §3) that Newton's laws do not determine the future motion of the mass. It may remain forever at the apex, or, at any time $t = T \geq 0$, it may spontaneously move in any direction Θ with the motion:

$$\begin{aligned} r_T(t) &= (1/144)\,(t-T)^4 \qquad t \geq T \\ &= 0 \qquad\qquad\qquad\quad t \leq T \\ \theta_\Theta(t) &= \Theta \end{aligned}$$

for Θ some constant angle. Each value of T and Θ yields a distinct motion compatible with the initial condition. The key property of the example is that Newtonian physics provides no probabilities for the different directions in which the spontaneous motion may proceed or for its timing. I have argued in Norton (2010), that it cannot provide such probabilities for the timing unless we artificially add further physical structure, such as a time constant.

As before, there are mappings under which Curie's principle fails and mappings under which it succeeds:

	Curie's principle fails	*Curie's principle succeeds*
cause	Mass-dome system at t=0.	Mass-dome system at t=0.
effect	A particular spontaneous motion, $(r_T(t), \theta_\Theta(t))$, for some specific T and Θ.	The set of all possible motions $M = \{(r_T(t), \theta_\Theta(t))$: all Θ, $T \geq 0\}$
rule of dependency	Newton's laws of motion.	Newton's laws of motion.

For the "fails" case, the cause, the mass-dome system at t=0, is symmetric under rotations of the dome around the apex. The effect, however, does not manifest this symmetry, since the

motion is always in some particular direction Θ.[8] This arises since Newton's laws turn out not to respect the rotational symmetry of this particular system (which is an unexpected outcome for those of us whose expectations are set by ordinary textbook treatments of Newtonian systems).

The change in the mapping that enables Curie's principle to succeed is to alter what is mapped as the effect. A particular spontaneous motion is replaced by the set of all possible motions M, as indicated. This is quite analogous to the shift of the effect from a particular radioactive decay event to the probability distribution to which it conforms. This set M does manifest the rotational symmetry of the cause. That is, if R_α is a rotation over the dome by angle α about the apex of the dome, then an individual spontaneous motion is mapped to a new one:

$$R_\alpha(r_T(t), \theta_\Theta(t)) = (r_T(t), \theta_{\Theta+\alpha}(t))$$

If $(r_T(t), \theta_\Theta(t))$ is in the set M, then so is $(r_T(t), \theta_{\Theta+\alpha}(t))$. Hence it follows that[9]

$$R_\alpha M = M$$

for all angles α. Thus Curie's principle succeeds. That Newton's laws do not respect rotational symmetry for individual spontaneous motions is no longer a problem. Newton's laws do respect this symmetry when applied to the set of all possible spontaneous motions, compatible with the cause.

4.3 Success and Failure for Type versus Token Causation

Once again it is tempting to protect the standard view that Curie's principle fails for indeterministic systems by favoring one mapping of cause and effect over another. We might argue that mapping the effect to the set of all motions yields success selectively by excising just that part of the effect that would lead to failure.

The temptation should be resisted. There is no unique, correct mapping of cause and effect into the examples. Both described here are admissible. They merely correspond to different senses of cause and effect. The distinction is so familiar that the senses have different

---

[8] The exception is the case in which the mass remains forever at the apex. We might imagine that case included in the formulae above as $T=\infty$.

[9] That is, $R_\alpha M = \{R_\alpha(r_T(t), \theta_\Theta(t))$: all $\Theta, T\geq 0\} = \{ (r_T(t), \theta_{\Theta+\alpha}(t))$: all $\Theta, T\geq 0\}$

$= \{ (r_T(t), \theta_{\Theta'}(t))$: all $\Theta', T\geq 0\} = M$, where $\Theta' = \Theta+\alpha$.

names. One is "type causation": treatment with penicillin cures bacterial infections. The other is "token causation": treatment of this particular patient on such and such days cured this particular patient's bacterial infection. The two senses can separate. At the type level, smoking causes lung cancer. But it is harder to maintain that causal relation at the token level when a majority of smokers do not contract lung cancer.

In the last two examples of radioactive decay and Newtonian indeterminism, Curie's principle succeeds for type causation and fails for token causation. The sense we select will match our purposes and perhaps whims. A smoker's concern is his or her own specific well-being. Such a smoker may concentrate on a failure of token causation, at least in the sense that this smoker's smoking will neither assuredly or even probably lead to the smoker contracting lung cancer. Public health officials will focus on type causation: in general, smoking causes lung cancer in the sense that it raises the average cancer rate in the population. They seek to advance the overall health of the population and, for them, the averages matter.

Correspondingly, might we argue that the initial state of the radioactive atom or the dome is not properly the cause of the specific decay or spontaneous motion, but rather only of the tendencies and possibilities encoded in the probability distribution $\rho(\theta, \phi, t)$ or set M. That view favors type causation and the success of Curie's principle. Both Chalmers (1970, p. 146) and Ismael (1997, §6) protect Curie's principle from failure in the case of radioactive decay in just this way.

5. Conclusion

Causal metaphysics is a troubled field. It is had no content beyond an elaborate exercise in naming, that is, the attaching of causal labels to pre-existing science. While evocative labeling can be conceptually helpful in so far as it aids us in forming apt mental pictures, mere labeling falls short of what causal metaphysics sometimes purports to offer: factual restrictions on all possible sciences in virtue of their causal characters.

Elsewhere (Norton, 2003, 2007, 2009), I have argued for the failure of efforts to locate such a factual principle of causality that usefully restricts our science. Curie's principle is another such failure. Symmetries of a crystal lattice and imposed fields must reappear as symmetries of the crystal's properties. Symmetries of an initial state, propagated by a symmetry

preserving rule of time evolution, must reappear as symmetries in the propagated state. However these are not instances of a more general, factual causal principle to which all science must conform. Whether the principle succeeds or fails, I have argued, is a matter of how we choose to attach causal labels to our science. This pliability of choice is what makes Curie's principle a pliable truth, that is, a truism; or at least it is in cases in which we deem it to succeed.

References

Brading, Katherine, and Castellani, Elena. eds. 2003. *Symmetries in Physics: Philosophical Reflections*. Cambridge: Cambridge University Press.

Brading, Katherine and Castellani, Elena. 2013. "Symmetry and Symmetry Breaking," The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/symmetry-breaking/>

Belot, Gordon. 2003. "Notes on Symmetries" in Brading and Castellani. 2003, pp.393-412.

Castellani, Elena. 2003. "On the Meaning of Symmetry Breaking." pp. 321-334 in Brading and Castellani. 2003.

Chalmers, Alan F., 1970. "Curie's Principle", *British Journal for the Philosophy of Science*, 21: 133–148.

Curie, P., 1894. "Sur la symétrie dans les phénomènes physiques, Symétrie d' un champ électrique et d'un champ magnétique." *Journal de Physique*, 3rd ser., 3: 393–417; in *Oeuvres de Pierre Curie*. Paris: Gauthier-Villars, pp. 118-141.

Earman, John. 2004. "Curie's Principle and Spontaneous Symmetry Breaking." *International Studies in the Philosophy of Science* 18: 173–198.

Ismael, Jenann. 1997. "Curie's Principle", *Synthese*, 110: 167–190.

Nakamura, Norihiro, and Nagahama, Hiroyuki, 2000. "Curie Symmetry Principle: Does it Constrain the Analysis of Structural Geology?" *Forma*. 15: 87-94.

Norton, John D. 2003. "Causation as Folk Science." Philosophers' Imprint Vol. 3, No. 4 http://www.philosophersimprint.org/003004/; reprinted in pp. 11-44 *Causation,*

Physics and the Constitution of Reality. H. Price and R. Corry, eds. Oxford: Oxford University Press.

Norton, John D. 2007. "Do the Causal Principles of Modern Physics Contradict Causal Anti-Fundamentalism?" pp. 222-34 in *Thinking about Causes: From Greek Philosophy to Modern Physics*. eds. P. K. Machamer and G. Wolters, Pittsburgh: University of Pittsburgh Press.

Norton, John D. 2009 "Is There an Independent Principle of Causality in Physics?" *British Journal for the Philosophy of Science*. 60: 475-86.

Norton, John D. 2010. "There are No Universal Rules for Induction." *Philosophy of Science*. 77: 765-77.

Roberts, Bryan W. 2013. "The Simple Failure of Curie's Principle." *Philosophy of Science*. 80: 579-592.

van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford: Clarendon.

How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE

Rune Nyrup, Durham University, rune.nyrup@dur.ac.uk

Abstract: This paper defends an account of explanatory reasoning generally, and inference to the best explanation in particular, according to which it first and foremost justifies pursuing hypotheses rather than accepting them as true. This side-steps the problem of why better explanations should be more likely to be true. I argue that this account faces no analogous problems. I propose an account of justification for pursuit and show how this provides a simple and straightforward connection between explanatoriness and justification for pursuit.

1. Introduction

Most proponents of *inference to the best explanation* (IBE) take it to be a distinctive mode of non-deductive inference where *explanatory reasoning*, i.e. considerations concerning what would be a good or the best explanation of one or more phenomena, is used as a guide to theory choice. This form of reasoning, they hold, is in general a reliable, if fallible, guide to the truth of hypotheses.

The idea of an explanatory inference goes back to Charles Peirce who promoted an inference, which he called *abduction*, proceeding from the premise that a given hypothesis, if it were true, would make an otherwise surprising fact "a matter of course" (CP 5.189).[1] Recent scholarship has however emphasised that Peirce's mature account of abduction differs significantly from the contemporary notion of IBE.[2] Contemporary discussions usually assume that explanatory reasoning, at least in the form of IBE, can justify *accepting* hypotheses as (approximately, partially, etc.) true. They thus regard it as a species of inductive or ampliative inference. This view is often called *Explanationism*.

1 All references to Peirce (1932-58) are abbreviated: CP [volume].[paragraph number].

2 The following interpretation is defended especially clearly by McKaughan (2008). Cf. also Hintikka (1998), Minnameier (2004), Paavola (2006), Campos (2011).

While Peirce agreed that abductions should guide our choices of hypotheses, he only understood this in the sense of choosing which hypotheses to investigate further. Peirce held that only empirical investigations can justify accepting a hypothesis, insisting that abduction gives us no reason to regard a hypothesis as true, except insofar as it leads to subsequent empirical testing which the hypothesis passes. He did regard abduction as a form of inference which involves giving reasons (whether good or bad) and not, for instance, a mere heuristic for "discovery". However, these are reasons for *courses of action*, viz. subjecting hypotheses to empirical testing, rather than reasons for belief or acceptance (McKaughan 2008, 450 & 454).

In this paper I defend a view of the justificatory role in science of explanatory reasoning in general, and IBE in particular, along the lines of these Peircean insights.[3] Specifically, drawing on the distinction between acceptance and pursuit (Laudan 1977; Franklin 1993a), I propose to see explanatory reasoning as first and foremost providing justification for *pursuing* a hypothesis, as opposed to justification for *accepting* it.

The Peircean view defended here avoids what Peter Lipton (2004) calls *Voltaire's Objection* to explanationism: why should we regard a hypothesis as any more likely to be *actually* true just because it would be a better explanation if it *were* true? The Peircean view side-steps this problem since it requires no general connection between explanatoriness and truth. Furthermore, the Peircean view faces no analogous problem either. As I shall show, there is a simple and straightforward connection between good explanations and justification for pursuit, based on the kinds of "economical" considerations Peirce stressed as fundamental to abduction. I introduce Voltaire's Objection in Section 2 and explain why it poses a problem to explanationism. In Section 3, I present a

3 I do not claim this view to be the most plausible interpretation of Peirce's considered views on abduction, much less to capture everything Peirce ever wrote about it. I merely use it as a name for the view, inspired by Peirce, which I defend in the context of the contemporary debate.

general account of pursuit and then, in Section 4, show how explanatory reasoning can justify pursuit.

2. Voltaire's Objection

The slogan that one should infer "the best" explanation conceals an important distinction. For there are at least two senses in which an explanation can be better than its competitors, and these should be kept separate when evaluating explanatory inferences (Lipton 2004, ch. 4). In the first case, a hypothesis may be the *likeliest* explanation relative to the other competing hypotheses considered. The likeliness of an explanation has to do with truth – it is the explanation which we regard as most probably true, or closest to the truth, etc. For instance, we may be able to rule out, or make highly improbable, all plausible alternative explanations in light of the available evidence and accepted background theories. Here, the remaining explanatory hypothesis would be the likeliest available explanation, and in this sense the best. As Lipton is careful to point out, IBE is only interesting as an inductive inference to the extent that it goes beyond merely being an inference to the likeliest explanation. Since scientists generally aim to discover good explanations, if a hypothesis *H* is the likeliest available explanation of some otherwise surprising phenomenon, they would be justified in accepting *H*. For my purposes here, this is a perfectly cogent inference and nothing I say in this paper aims to challenge it.

The sense of "best explanation" that is of interest to explanationists concerns how good an explanation we would deem a hypothesis *H* to be, if it were true. Let us say that the *explanatoriness* of *H* is the amount and quality of the explanations *H* would provide, if it were true. Or, since the "goodness" of explanations is usually taken to concern how much understanding they give us, the explanatoriness of *H* can also be understood as the amount of understanding *H* could potentially afford us.[4] Assessing this requires subjunctive

4 Explanatoriness is my preferred term for what Lipton calls "loveliness".

reasoning, i.e. reasoning about what would be the case – viz. how much understanding it would provide – if *H* were true. We can call this kind of reasoning *explanatory reasoning*. What explanationists claim, then, is that explanatory reasoning can give us additional or independent reason to accept a hypothesis as true (or approximately true). In other words, they regard the explanatoriness of a hypothesis as a guide to its likeliness.

This claim is also what makes the explanationist account of IBE controversial. One question concerns what "good explanations" means. There are many different accounts of explanation (causal, unification, etc.), and these variably emphasise certain virtues (being simple, unifying, coherent, elegant, quantitatively precise, specifying a mechanism, etc.) as characteristic of good explanations. Since my argument in this paper does not depend on any particular view of explanation or of how they give us understanding (however we conceive of this), I stay neutral on these matters.

Explanationism however faces a more pressing problem – what Lipton (2004, ch. 9) calls *Voltaire's Objection*. As critics have pointed out, the fact that a hypothesis would be a good explanation of something, if it *were* true, does not, *prima facie*, seem to have any implications for whether it is *actually* true. Indeed, this seems worryingly close to a form of wishful thinking. So why should this give us any additional reason to accept the hypothesis?[5] Of course, like all inductive inferences, IBE would be fallible, and so explanationists should not be expected to *guarantee* its success. Nonetheless, they ought to provide some reason to think that explanatoriness is generally a reliable guide to likeliness or that it generally tends to take us closer to the truth.

My focus in this paper is however not on the arguments explanationists give for the reliability of IBE.[6] Rather, I restrict myself to showing that the Peircean view avoids Voltaire's Objection altogether and, furthermore, faces no analogous problems.

5 See Barnes (1995) for a sustained criticism along these lines directed specifically at Lipton.

6 See Douven (2011, sec. 3.2) for a brief overview.

3. Pursuing Hypotheses and Justifying It

In his exegetical study of Peirce's views on abduction, Daniel McKaughan (2008) distinguishes three general interpretations: the *Generative Interpretation*, the *Justificatory Interpretation*, and the *Pursuitworthiness Interpretation*. The Justificatory Interpretation corresponds to explanationism, where abduction is taken to provide justification for accepting hypotheses as true or approximately true. This view is typically contrasted with the Generative Interpretation, associated with Hanson (1958) (e.g. Paavola 2006). Hanson argued that it is a significant philosophical task to analyse the processes through which scientific theories are formulated, generated or discovered, promoting Peirce's abduction as such an analysis. Popper (1934/1959) and the positivists, he argued, were mistaken in restricting philosophy to questions of how evidence justifies the acceptance of theories, relegating all other issues to empirical sociology, psychology or history.

McKaughan argues that these two interpretations overlook an important step in the process of inquiry between the initial formulation of a hypothesis and its subsequent acceptance (or rejection) as part of established scientific knowledge. Apart from formulating and developing hypotheses to investigate, scientists, in order to prioritise their time, resources, and efforts, furthermore need to make decisions regarding which of these to investigate or develop further. In other words, scientists need to make decisions regarding which hypotheses are most worthy of further *pursuit*.[7] As McKaughan shows, this was a dominant theme especially in Peirce's later discussions of abduction – thus, the Pursuitworthiness Interpretation. It is this aspect of Peirce's views on which I draw in the following.

7 Pursuing a hypothesis is generally taken to involve at least two aspects: (i) subjecting it to empirical testing and (ii) developing it theoretically, e.g. clarifying it, resolving conceptual problems, or removing apparent tensions with other accepted theories (Laudan 1977; Whitt 1990). I focus on (i) in this paper.

The distinction between *accepting* and *pursuing* a hypothesis was first coined (in those terms) by Larry Laudan (1977, 108-14; 1980, 174). Laudan noticed that, historically, scientists have often chosen to work on scientific theories despite these having have major empirical and conceptual problems relative to the dominant views, citing, amongst others, Copernicanism, the atomic theory, and quantum mechanics in their early stages. By distinguishing between pursuing and accepting, Laudan argued, we can say that it was rational for scientists to pursue these theories even though there where strong reasons to accept competing theories. More recently, Allan Franklin (1993a, 1993b) argues that certain episodes in particle physics are best understood as cases where physicists chose to pursue hypotheses before they had reasons to accept them.

Franklin's case studies are especially suggestive for present purposes, since these concern hypotheses that were pursued exactly because of their potential for explaining otherwise puzzling phenomena. For example, Franklin (1986, ch. 1) discusses the rejection by particle physicists of the so-called principle of parity conversation. The puzzling phenomenon physicists faced was this: for a specific set of decay patterns, the principle that each particle has a unique mass indicated that these decays stemmed from a single particle, while the principle of parity conversation ruled this out. When the physicists T.D. Lee and C.N. Yang in 1956 proposed that parity conversation may be violated in weak interactions, and suggested experiments to test this hypothesis, it sparked an intense experimental interest. It should be noted, first, that the same hypothesis had earlier been suggested as a logical possibility, but without being proposed as a solution to the above puzzle and without arousing much interest (Franklin 1986, 29f). Second, many of the physicists involved were quite convinced that the experiments would falsify the hypothesis.[8]

8 Franklin reports (1986, 24) that Richard Feynman bet Norman Ramsey $50 to $1 that the experiments would fail to show parity violation – and ended up paying!

Apart from the descriptive point that scientist often actually do make and argue for decisions about which hypotheses to pursue, there are also normative reasons why scientists ought to justify such choices.[9] The reason is pragmatic: the resources available to scientists are scarce but human imagination is abundant. In Peirce's words:

> Proposals for hypotheses inundate us in an overwhelming flood, while the process of verification to which each one must be subjected before it can count as at all an item, even of likely knowledge, is so very costly in time, energy, and money—and consequently in ideas which might have been had for that time, energy, and money, that Economy would override every other consideration even if there were any other serious considerations. In fact there are no others. For abduction commits us to nothing. It merely causes a hypothesis to be set down upon the docket of cases to be tried (CP 5.602)[10]

In other words, scientists need to justify which hypotheses are worth investigating in order to prioritise their resources. Justifying pursuit is, essentially, a decision-theoretic problem of how to optimise the epistemic output of science.

Although justification for pursuit is motivated by practical or pragmatic issues, it is not wholly detached from epistemic matters. On the contrary, it is still concerned with how to best or most effectively achieve our epistemic goals. This also makes it slightly misleading to characterise the distinction as one between justification and pursuit. Although the two are sometimes conflated, the distinction between (justification for) accepting and pursuing hypotheses cuts across the much discussed distinction between

9 Further case studies of pursuit are discussed by Whitt (1990) and McKaughan (2008).

10 Peirce frequently connects “economical” considerations to his account of abduction. See McKaughan (2008, 452ff) for further references.

context of discovery/context of justification.[11] Choices regarding which hypotheses to accept as well as which to pursue can and ought to be justified. The difference is that acceptance concerns which hypotheses are more likely to be to true, given our background knowledge and evidence, whereas justification for pursuing hypotheses involves practical reasoning about which *courses of action to follow*, given our resources, overall goals, and available information.[12]

How do we decide which hypotheses we are justified in pursuing, then? To answer this we must first, as Šešelja, Kosolosky & Straßer (2012) point out, make clear what kinds of goals we are justifying pursuit relative to. If we are interested in a broader set of moral, political and epistemic goals (as e.g. Kitcher (2011)) we need to take things like ethical implications and technological progress into account. In this paper I am however focusing only on epistemic or intellectual goals such as learning the truth or getting better explanations or understanding. This focus also seems to be assumed by explanationists – ethical implications or potential technological applications are usually taken to be irrelevant to the explanatoriness of a hypothesis.

Given this focus, we are justified in pursuing that course of action we judge will bring us the closest to achieving our epistemic goals. Doing will typically involve, as McKaughan (2008) points out, somehow *weighing* and *ranking* the salient competing hypotheses in terms of factors we take to be relevant to determining this. What these factors are exactly will presumably vary somewhat from case to case, but some general suggestions can be made. Thus, Peirce highlights the "cost, the value of the thing proposed, in itself; and its effect upon other projects" (CP 7.220). Elaborating on Peirce, McKaughan

11 Laudan (1980, 174) characterises context of pursuit as a "nether region" between discovery/generation and (ultimate) justification. In my view, the "context" terminology is still misleading: these are not neatly separated phases or contexts of scientific inquiry. The distinction concerns different kinds of *choices*, which may overlap, and the kinds of *justification* relevant to them.

12 McKaughan (2008, 454); cf. Kapitan (1992).

mentions "our time, resources, and value of the estimated payoff in comparison to other courses of action … If we estimate that testing the hypothesis will be *easy*, of potential *interest*, and *informative*, then we should give it a high priority" (2008, 457). Independently, Franklin (1993a, 122) observes from his case studies that "[t]he decision to pursue an investigation seems to depend on a weighting of at least three factors; the interest of the hypothesis, its plausibility, and its ease of test". He also mentions (1993b) more pragmatic concerns such as "recycling expertise" or being able to continue already ongoing research programmes.

This of course raises the question of how these factors should be weighed against each other. In practice, this will probably be a matter of informed judgement. But in order to clarify the underlying logic, it can be useful to think of it terms of decision-theoretic models of simplified or idealised situations. To illustrate this, I will in the following develop a model that is particularly useful for thinking about explanatory reasoning.

This model focuses on just three types of outcomes of pursuing a hypothesis H:

i. We get strong enough evidence in favour of H to accept it.
ii. We get strong enough evidence against H to reject it.
iii. We get inconclusive evidence, and so stay agnostic.

We can abbreviate each of these outcomes as $a(H)$, $r(H)$ and $\sim a(H)\&\sim r(H)$, respectively. So we are ignoring how to figure in the costs of pursuing H, whether pursuing H might reveal other interesting things about the world, as well the possible "effects upon other projects" or Franklin's pragmatic factors.

Let $EV(a(H))$, $EV(r(H))$ and $EV(\sim a(\mathrm{H}) \,\&\, \sim r(H))$ represent the *epistemic value* associated with each of the three outcomes obtaining. We can think of this as the degree to

which each of these outcomes would take us towards or away from reaching our epistemic goals. It corresponds roughly to what Peirce, McKaughan and Franklin call the "value" or "interest" of the hypothesis. Since pursuing *H* has a causal influence on which of outcome obtains, we should weigh each of these in terms of how probable they are to obtain *given* that we pursue it. Let $p(H)$ be the decision to pursue *H* and let $EEV(H)$ be the expected epistemic value of pursuing *H*, we thus have:[13]

$$\begin{aligned} (1) \qquad EEV(p(H)) = \; & EV(a(H)) * Pr(a(H) \mid p(H)) \\ + \; & EV(r(H)) * Pr(r(H) \mid p(H)) \\ + \; & EV(\sim a(\mathrm{H}) \,\&\, \sim r(H)) * Pr(\sim a(H) \,\&\, \sim r(H) \mid p(\mathrm{H})) \end{aligned}$$

Since we are ignoring the costs and other effects of pursuing *H*, it is natural to stipulate for simplicity that the value of staying agnostic is nil, and so drop the last line.

Now, how epistemically valuable it would be to accept *H*, and how likely we are to get evidence for or against it, presumably depends on whether *H* is actually true. To make this explicit in the model, we can conditionalise on the truth and the falsity in each line, giving us:

$$\begin{aligned} (2) \qquad EEV(p(H)) = \; & EV(a(H) \,\&\, H) * Pr(a(H) \mid H \,\&\, p(H)) * Pr(H) \\ + \; & EV(a(H) \,\&\, \sim H) * Pr(a(H) \mid \sim H \,\&\, p(H)) * Pr(\sim H) \\ + \; & EV(r(H) \,\&\, H) * Pr(r(H) \mid H \,\&\, p(H)) * Pr(H) \\ + \; & EV(r(H) \,\&\, \sim H) * Pr(r(\mathrm{H}) \mid \sim \mathrm{H} \,\&\, p(\mathrm{H})) * Pr(\sim \mathrm{H}) \end{aligned}$$

13 The probabilities can be interpreted either as objective chances or credences, depending on whether one is interested in externalist or internalist justification for pursuit. In the latter case, the conditional probabilities should be interpreted as the credence that pursuing *H* will bring about the outcome.

In this model, then, we would be justified in pursuing that hypothesis *H* which maximises *EEV*(*p*(*H*)).[14]

One attractive feature of this model is that it explicitly represents a number of the factors mentioned earlier, and furthermore calls attention to some factors left out. I have already mentioned that *EV*(*a*(*H*) & *H*) and *EV*(*r*(*H*) & ~*H*) represents how valuable or interesting it would be to know whether *H* is true. Correspondingly, *EV*(*a*(*H*) & ~*H*) and *EV*(*r*(*H*) & *H*) is how problematic it would be to mistakenly accept a falsehood or reject a truth. The unconditional probabilities represent how likely or plausible *H* (and ~*H*) is prior to testing; and the conditional probabilities represent how likely we are to get reliable and misleading evidence, respectively.

Models of this kind are of course both idealised and abstract. I do not suppose that it is generally possible to make anything but rough estimates or comparisons of these factors. Furthermore, the estimates of individual scientists, as well as what they take the most important epistemic outcomes of science to be, probably varies significantly. I do not have any comprehensive account of these matters. Finally, scientists obviously do not always conform to or even approximate this model in their deliberations about which hypotheses to pursue even when their goals are purely epistemic; nor do I claim that it would be better if they did. Nonetheless, I find that this kind of models provides a useful normative framework for expressing and clarifying issues regarding justification for pursuit. In the following I apply it to the case of explanatory reasoning.

4. How Explanatory Reasoning Justifies Pursuit

I claim that the Peircean view avoids Voltaire's Objection. In a nutshell, I claim that explanatory reasoning justifies pursuing a hypothesis *H* by showing that it would be more

14 The model becomes more complicated if we take into account possible synergy effects of pursuing more than one hypothesis simultaneously.

epistemically valuable to learn that *H* is true than its salient competitors.

To spell out this argument in more detail, notice first that the epistemic goals of science include more than simply knowing as many truths as possible. As Philip Kitcher (1993, 94) puts the point:

> Tacking truths together is something any hack can do. … The trouble is that most of the truths that can be acquired in these ways are boring. Nobody is interested in the minutiae of the shapes and colors of the objects in your vicinity, the temperature fluctuations in your microenvironment, the infinite number of disjunctions you can generate with your favorite true statement as one disjunct, or the probabilities of the events in the many chance setups you can contrive with objects in your vicinity. What we want is *significant* truth.

There are plenty of trivial truths out there that could be discovered and at much lower cost than the questions actually pursued by scientists. The value of scientific knowledge depends on other factors beyond merely the amount of truths known, no matter how certain.

Now, what these additional factors are – what other "epistemic goods", as we might call them, are important in science – is not something we need to give a general account of here. However, most philosophers of science, and explanationists in particular, seem to agree that having good explanations is among them.[15] So one way a hypothesis can be more epistemically valuable than merely being true is by being a good explanation, i.e. by increasing our understanding of the phenomena scientists investigate. Philosophers may

15 For instance, Kitcher (1993, 105ff) discusses "Explanatory Progress" as one of the goals pursued by science beyond mere truth.

disagree about why explanation and understanding are epistemically valuable – maybe they are intrinsically valuable, or maybe they are only valuable as a means to achieving other important epistemic goals. However, all I need for the present argument is that explanation/understanding is in fact epistemically valuable.

Now, consider the premise of an IBE: that the hypothesis *H* would provide the most understanding out of a set of rival explanations, if it were true. Thus, if we were to learn that *H* is actually true, this would be an epistemically valuable outcome. Indeed, learning that the most explanatory hypothesis is true would be the optimal epistemic outcome as far as explanation and understanding are concerned. Suppose, then, that everything else is held equal between a set of rival hypotheses: the costs of pursuing them are the same, we regard it as equally likely that pursuing them would give us reliable evidence for or against them, all other expected epistemic outcomes of pursuing them are equal, and so on. In this case, given the account of justification for pursuit outlined above, scientists would be justified in pursuing the most explanatory hypothesis (given that we focus on epistemic goals).

To express this in terms of the decision-theoretic model developed earlier, we can express the assumption that explanatoriness is *one* important epistemic goal as the claim that if H_1 is more explanatory than H_2, then, all else being equal, $EV(a(H_1)\ \&\ H_1) > EV(a(H_2)\ \&\ H_2)$.[16] Notice furthermore, from equation (2), if $EV(a(H_1)\ \&\ H_1) > EV(a(H_2)\ \&\ H_2)$ then, all else being equal, $EEV(p(H_1) > EEV(p(H_2)$. So it follows that if H_1 is more explanatory than H_2, we are, all else being equal, justified in pursuing H_1 rather than H_2.

The argument can be illustrated by an analogy: Suppose a team of treasure hunters know of two caves, C_1 and C_2, where a large treasure might be stashed. As far as they know the treasure is equally likely to be in either cave, but they only have the resources to send an expedition to one of them. However, they do know that C_1 could hold up to twice the

16 This is "all else being equal" since H_2 might be more valuable with regards to other epistemic goals besides explanatoriness.

amount of treasure that C_2 could. Assume that this does not give them any further information about its location or how difficult or expensive it would be to recover. Nonetheless, it is still more rational (for obvious decision theoretic reasons) for them to send the expedition to explore C_1 rather than C_2.

This argument shows that IBE can justify pursuit, all else being equal. In other words, explanatoriness can serve as a tie-breaker to justify pursuing one hypothesis rather than certain others. More generally, it should also be clear that if a hypothesis has a high degree of explanatoriness this adds to the expected epistemic value of pursuing it and thus gives *some* additional reason to pursue it, although not always a *decisive* reason.

Notice that I am assuming that we are deciding which hypothesis to pursue after we have fixed our estimates of all factors relevant to pursuit. If we, for instance, *discover* that a hypothesis is more unifying than we previously thought, or *change* it to become more unifying, this can influence our estimates of the other factors. So if revising the hypothesis makes it less plausible, this might cancel out any gains in explanatoriness. Similarly, we had to assume in the treasure hunter analogy that knowing the size of the cave does not provide additional information about the location of the treasure, or that they have already taken this into account.

Since nothing in this argument assumes a connection between explanatoriness and likeliness, this shows why the Peircean View avoids Voltaire's Objection. Let me close by considering a possible objection: Justifying the pursuit of a hypothesis still involves showing it to be minimally plausible or probable. Indeed, Peirce sometimes says that abductions give us "reason to suspect that [the hypothesis] is true" (CP 5.189) or reasons "regarded as lending the hypothesis some plausibility" (CP 2.511, footnote) and that "[c]ertain premises will render an hypothesis probable, so that there is such a thing as legitimate hypothetic inference [i.e. abduction]" (*loc. cit.*). However, if this is the case, the

Peircean view would also require *some* connection between explanatoriness and likeliness (or plausibility), even if a weaker one than explanationists tend to require. But this is sufficient for a version of Voltaire's Objection to apply to the Peircean view as well.

The premise of this objection is mistaken. Justification for pursuit need not stem from showing the hypothesis any more probable or plausible than before. Even if a necessary condition for a hypothesis being pursuitworthy is some minimal degree of plausibility, it is not sufficient. *One* way of justifying pursuit might be to show that the hypothesis is more plausible than previously thought. However, this is not the only way. For one thing, one could equally argue that a hypothesis is only worth investigating if it is not completely trivial or obvious.[17] Thus, could also justify pursuing a hypothesis by showing that there is more reason to doubt it than previously thought. And, as argued above, justification for pursuing a hypothesis can also stem from how interesting or valuable it would be to know whether it is true, independently of its plausibility.

Furthermore, it is not generally the case that having higher plausibility gives us more reason to pursue a hypothesis. Consider equation (2) again. If $Pr(H_1) > Pr(H_2)$ it does not follow that, all else being equal, $EEV(H_1) > EEV(H_2)$. First, raising $Pr(H_1)$ gives more weight to both the first and the third term in equation (1). So if, say, $EV(a(H_1), H_1) * Pr(a(H_1) \mid H \,\&\, p(H_1)) < EV(r(H_1) \,\&\, H_1) * Pr(r(H_1) \mid H_1, p(H_1))$ – which by assumption is the same for H_1 and H_2 – this make $EEV(H_1)$ *lower* than $EEV(H_2)$. Second, raising $Pr(H_1)$ at the same time lowers $Pr(\sim H_1)$, thus lowering the second and the third term. Again, depending on our estimates of the other factors, this could lower $EEV(H_1)$.

In sum, although being very likely or plausible can *sometimes* be a good reason to pursue a hypothesis, we can equally be justified in pursuing a hypothesis exactly *because*

17 In fact, neither of these conditions are necessary. As Franklin (1993a, ch. 3) points out, physicists sometimes pursue experimental work on a hypothesis after they regard it as conclusively falsified. Pursuing *H* can serve other epistemic goals beyond merely generating evidence for or against *H*.

we think it very likely false and it would be easy to show this. And this was in fact something Peirce often stressed:

> the best hypothesis … is the one which can be the most readily refuted if it is false. This far outweighs the trifling merit of being likely (CP 1.120)

This is also a plausible interpretation of why the physicists in Franklin's (1986) story chose to pursue the parity violation hypothesis, despite thinking it very likely to be false.

5. Conclusion

The argument given in this paper is quite general. It only rests on the premise that it, all else being equal, is more epistemically valuable to know whether more explanatory hypotheses are true than less explanatory ones. In particular, I have not presupposed any specific account of explanation or of why explanations are valuable. Combined with the account of justification for pursuit outlined in section 2, I have shown how this gives us a simple and straightforward connection between explanatoriness and justification for pursuit. The Peircean view avoids Voltaire's Objection and faces no analogous problems.

References:

Barnes, Eric. 1995. "Inference to the Loveliest Explanation." *Synthese* 103:251-277.

Campos, Daniel. 2011. "On the Distinction Between Peirce's Abduction and Lipton's Inference to the Best Explanation." *Synthese* 180:419-42.

Douven, Igor. 2011. "Abduction". In *The Stanford Encyclopedia of Philosophy.* Spring 2011 Edition, ed. Edward Zalta. URL = <http://plato.stanford.edu/archives/spr2011/entries/abduction/>.

Franklin, Allan. 1986. *The Neglect of Experiment*. Cambridge: Cambridge University Press.

—— 1993a. *The Rise and Fall of the Fifth Force: Discovery, Pursuit, and Justification in Modern Physics.* New York: American Institute of Physics.

—— 1993b. "Discovery, Pursuit, and Justification." *Perspectives on Science* 1:252–84.

Hanson, Norwood. 1958. *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.

Hintakka, Jaakko. 1998. "What Is Abduction? The Fundamental Problem of Contemporary Epistemology." *Transactions of the Charles S. Peirce Society* 34:503-33.

Kapitan, Tomis. 1992. "Peirce and the Autonomy of Abductive Reasoning." *Erkenntnis* 37:1-26.

Kitcher, Philip. 1993. *The Advancement of Science: Science without Legend, Objectivity without Illusions.* Oxford: Oxford University Press.

—— 2011. *Science in a Democratic Society*. Amherst, N.Y.: Prometheus.

Laudan, Larry. 1977. *Progress and Its Problems: Towards a Theory of Scientific Growth.* London: Routledge.

—— 1980. "Why Was the Logic of Discovery Abandoned?" In Nickles, T. (ed.) 1980. *Scientific Discovery, Logic and Rationality*. Boston Studies in the Philosophy of

Science 56. Dordrecht: Reidel, pp. 173-83.

Lipton, Peter. 2004. *Inference to the Best Explanation*. 2nd Edition. London: Routledge.

McKaughan, Daniel. 2008. "From Ugly Duckling to Swan: C. S. Peirce, Abduction, and the Pursuit of Scientific Theories." *Transactions of the Charles S. Peirce Society* 44:446-68.

Minnameier, Gerhard. 2004. "Why Inference to the Best Explanation and Abduction Ought Not to Be Confused." *Erkenntnis* 60:75-105.

Niiniluoto, Ilkka. 1999. "Defending Abduction." *Philosophy of Science* 66 (Proceedings): S436-S451.

Paavola, Sami. 2006. "Hansonian and Harmanian Abduction as Models of Discovery." *International Studies in the Philosophy of Science* 20:93-108.

Peirce, Charles. 1932–1958. *Collected papers of Charles Sanders Peirce.* Vols. 1–8. Weiss, Paul, Charles Hartshorne, and Arthur Burks, eds. Cambridge, MA: Harvard University Press.

Popper, Karl. 1934/1959. *The Logic of Scientific Discovery*. Repr. New York: Routledge.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Šešelja, Dunja, Laszlo Kosolosky, and Christian Straßer. 2012. "The Rationality of Scientific Reasoning in the Context of Pursuit: Drawing Appropriate Distinctions." *Philosophica* 86:51-82.

Whitt, Laurie. 1990. "Theory Pursuit: Between Discovery and Acceptance." In *Proceedings of the Biennial Meetings of the Philosophy of Science Association,* Vol. 1990, Volume One: Contributed Papers, Fine, Arthur, Micky Forbes, and Linda Wessels, 467–83. East Lansing: The Association.

Conditionalization and Credal Conservatism

Ilho Park (Chonbuk National University, South Korea)

Abstract

This paper is intended to show an epistemic trait of the Bayesian updating rule. In particular, I will show in this paper that Simple/Jeffrey/Adams Conditionalization is equivalent to what I call Credal Conservatism, which says that when we undergo a course of experience, our credences irrelevant to the experience should remain the same.

1 Introduction

Quine (1951) suggests epistemic conservatism as one of the factors that should be considered when accommodating new experience. According to this conservatism, our old beliefs should be conserved as far as possible when new experience is incorporated into our belief system. A similar thing goes with our credences. In particular, it seems to be another conservative norm that when we undergo a course of experience, our credences irrelevant to the experience should remain the same. I will call this norm *Credal Conservatism*. The modifier 'credal' is intended to express that the conservatism in question is related to credences that are often regarded as coherent degrees of belief—i.e. degrees of belief that conform to the probability calculus.

The philosophical position that degrees of belief should conform to the probability calculus is often called Bayesianism. In addition to probabilistic coherence, Bayesianism provides another epistemic rule governing how to update our credences after we undergo some experience: Conditionalization. Roughly speaking, this Bayesian belief updating rule requires us to update our credences by conditionalizing on the experience. Then, is there any connection between Bayesianism and Credal Conservatism? In this paper, I will

attempt to give an answer to this question. Especially, I will prove, in what follows, that the conservatism at issue, if formulated properly, is equivalent to Conditionalization.

This paper is structured as follow: The next section will be dedicated to introducing several versions of Conditionalization and their equivalent formulations. In Section 3, I will formulate, in a strict but intuitive way, Credal Conservatism using the concept of probabilistic independence and its variants. And I will prove in Section 4 that Credal Conservatism formulated so is equivalent to Conditionalization.

2 Varieties of Conditionalization

As was mentioned, Conditionalization is a rule that requires us to update our credences by conditionalizing on experience that directly changes some parts of our credal state. There are at least three types of experience that triggers belief updating by Conditionalization. These types are characterized according to what credences are directly changed by experience. The first type of experience directly changes an agent's credence in a particular proposition to 1 and nothing else. I will call this type 'Simple experience'. In particular, when a course of experience directly changes an agent's credence in E to 1 and nothing else, I will call this experience 'Simple experience regarding E'. The modifier 'Simple' is intended to express that this type of experience triggers belief updating by 'Simple Conditionalization'. Here is the simple version of Conditionalization:[1]

> **Simple Contionalization, SC:** When an agent undergoes simple experience regarding E,
>
> (S1) for all X, $q(X) = p(X|E)$.

Here p and q are the agent's old and new credence function, respectively. Her old credence function is one that she had before undergoing the experience, and her new credence function is one that she will have after undergoing the experience. (Here and below, p and q always refer to the relevant agent's old and new credence functions, respectively.)

[1] In the cases where there is no confusion, I will omit the proviso that conditional credences are well-defined. My discussion that follows does not depend on this assumption.

The second type of experience directly changes an agent's credence distribution over a particular partition and nothing else, and so triggers belief updating by what is often called 'Jeffrey Conditionalization'. Thus, I will call the second type of experience 'Jeffrey experience'. When a course of experience directly changes a credence distribution over $\mathbb{E}$ and nothing else, I will call this experience 'Jeffrey experience regarding $\mathbb{E}$'. (Here and below, $\mathbb{E}$ always refers to a partition $\{E_1, \cdots E_n\}$ whose members are mutually exclusive and collectively exhaustive.) Then, Jeffrey Conditionalization is formulated as follows:

> **Jeffrey Conditionalization, JC:** When an agent undergoes Jeffrey experience regarding $\mathbb{E}$,
>
> (J1) for all X, $q(X) = \Sigma_i q(E_i)p(X|E_i)$.

The third type of experience will be called 'Adams experience (regarding $\mathbb{E}$-given-F)'. This experience directly changes an agent's conditional credence distribution over $\mathbb{E}$-given-F and nothing else. Here 'changing an agent's conditional credence distribution over $\mathbb{E}$-given-F' means that for some $E_i \in \mathbb{E}$, an agent's conditional credence in E_i given F is changed. Like Simple and Jeffrey experience, this experience triggers belief updating by a version of Conditionalization, which is dubbed 'Adams Conditionalization'. This version is formulated as follows:

> **Adams Conditionalization, AC:** When an agent undergoes Adams experience regarding $\mathbb{E}$-given-F,
>
> (A1) for all X, $q(X) = \Sigma_i q(E_i|F)p(X|E_iF)p(F) + p(X\bar{F})$

Unlike SC and JC, AC was only recently proposed by Bradley (2005).[2]

Now, consider the relationship among each version of Conditionalization. Simple experience is a special case of Jeffrey/Adams experience. Consider an agent who undergoes Simple experience regarding E. We can say equivalently that the agent undergoes Jeffrey

[2]Some authors call SC 'Conditionalization' and JC 'Probability Kinematics'. Indeed, there are various names that refer to the first version of conditionalization—e.g. Bayesian Conditionalization, Strict Conditionalization, Classical Conditionalization and so forth. It was Bradley (2005) who named the third version of conditionalization 'Adams Conditionalization'. A similar discussion can be found in Wagner (2003).

experience regarding $\{E, \bar{E}\}$ so that her new credence in E is 1. And (J1) is equivalently transformed into (S1) when an agent undergoes Simple experience regarding E.[3] Likewise, we can also say equivalently that the agent undergoes Adams experience regarding $\{E, \bar{E}\}$-given-$\mathbf{T}$ so that the new credence in E given $\mathbf{T}$ is 1. (Here and below, $\mathbf{T}$ refers to a tautology.) And (A1) is equivalently transformed into (S1) when an agent undergoes Simple experience regarding E..[4] Thus, we can say that:

> **Proposition 1:** When an agent undergoes Simple experience regarding E, (J1) and (A1) each are equivalent to (S1).

In other words, SC is a special case of JC and AC. A similar conclusion can be drawn with regard to the relationship between JC and AC. We can show with no difficulty that:[5]

> **Proposition 2:** When an agent undergoes Jeffrey experience regarding $\mathbb{E}$, (A1) is equivalent to (J1).

That is, JC is a special case of AC.

For the discussion that follows, we also need to note some equivalent formulations of SC, JC and AC. Consider the following formulations:

(S2) $q(E) = 1$, and for all X, $p(X|E) = q(X|E)$.

(J2) For all X and $E_i (\in \mathbb{E})$, $p(X|E_i) = q(X|E_i)$.

(A2) $p(F) = q(F)$, and for all X and $E_i (\in \mathbb{E})$, $p(X|\bar{F}) = q(X|\bar{F})$ and $p(X|E_iF) = q(X|E_iF)$.

[3] Suppose that an agent undergoes Jeffrey experience regarding $\{E, \bar{E}\}$ so that the new credence in E is 1. According to JC, we have that for all X, $q(X) = q(E)p(X|E) + q(\bar{E})p(X|\bar{E})$. Note that the new credence in E is 1, that is $q(E) = 1$. Thus, we have that for all X, $q(X) = p(X|E)$.

[4] Suppose that an agent undergoes Adams experience regarding $\{E, \bar{E}\}$-*given*-$\mathbf{T}$ so that the new credence in E *given* $\mathbf{T}$ is 1. According to AC, we have that for all X, $q(X) = q(E|\mathbf{T})p(X|E\mathbf{T})p(\mathbf{T}) + q(\bar{E}|\mathbf{T})p(X|\bar{E}\mathbf{T})p(\mathbf{T}) + p(X\bar{\mathbf{T}})$. Note that the new conditional credence in E-*given*-$\mathbf{T}$ is 1, that is $q(E|\mathbf{T}) = 1$. Thus, we have that for all X, $q(X) = p(X|E)$.

[5] Consider a course of experience that directly changes an agent's credence distribution over $\mathbb{E}$ and nothing else. We can say equivalently that the experience in question directly changes an agent's credence distribution over $\mathbb{E}$-*given*-$\mathbf{T}$ and nothing else. According to AC, then, we have that for all X, $q(X) = \Sigma_i q(E_i|\mathbf{T})p(X|E_i\mathbf{T})p(\mathbf{T}) + p(X\bar{\mathbf{T}}) = \Sigma_i q(E_i)p(X|E_i)$.

It is well known that (S1), (J1) and (A1) are equivalent to (S2), (J2) and (A2), respectively.[6] Here we need to consider (J2), which is often called *Rigidity* or *Sufficiency*. With the help of (J2), we easily ascertain that JC requires that when an agent undergoes Jeffrey experience regarding $\mathbb{E}$, her conditional credences given each member of $\mathbb{E}$ should remain the same. Note that (S2) and (A2) also involve this kind of rigidity. Some authors like Christensen (1994, p.70) think, thus, that Rigidity itself shows a conservative spirit of Conditionalization. However, it is not clear how Rigidity itself is related to the aforementioned credal conservatism. Moreover, Rigidity poses somewhat difficult questions such as: Why should rational agents hold fixed the relevant conditional credences rather than other credences?[7] This kind of question is raised since it is unclear how Rigidity is related to our epistemological norms like consistency, truth-conduciveness, simplicity, and so forth. Thus, if we can expose clearly how Rigidity is related to Credal Conservatism that seems to be one of the epistemological norms, then we may find a clue for answering that question. For this purpose, we should first formulate explicitly Credal Conservatism.

3 Formulations of Credal Conservatism

Credal Conservatism says that when we undergo a course of experience, our credences irrelevant to the experience should remain the same. This conservatism should include conditional credences as well as unconditional ones. So, the conservatism can be re-described as follows:

> **CC:** When experience directly changes an agent's particular credences and nothing else, for all X and Y, $p(X|Y) = q(X|Y)$ if the experience is irrelevant to her conditional credence in X given Y.

Admittedly, CC is far from clear. In particular, in order to examine the relationship between Conditionalization and CC, we should first define the *irrelevance of conditional credences to*

[6] The relevant proofs are found in various papers or textbooks. For example, see Bradley (2005), Jeffrey (1992, 2004), and Howson and Urbach (1993).

[7] This question can be found in some introductory works on Bayesian epistemology—for example, see Talbott (2008). And there are discussions about why rational agents should hold fixed the conditional credences in question rather than others credences—for example, see Harper and Kyburg (1968) and Levi (1969).

Simple/Jeffrey/Adams experience. In what follows, I will consider several cases in which an agent undergoes such kinds of experience and then define probabilistically the irrelevance in question.

First, let's consider Simple experience that makes an agent update her credences by SC on E. How can we formulate probabilistically the irrelevance of a conditional credence in X given Y to this experience? Many Bayesians may define the irrelevance by means of the probabilistic conditional independence between X and E given Y. That is,

> **Definition 1:** Simple experience regarding E is *irrelevant to an agent's conditional credence in X given Y* if and only if $p(XE|Y) = p(X|Y)p(E|Y)$.

When $p(E|Y)>0$, the proposition that $p(XE|Y) = p(X|Y)p(E|Y)$ is equivalent to the proposition that:

(3.1) $\quad p(X|YE) = p(X|Y)$.

This says that the agent's conditional credence in X given Y remains the same no matter whether or not E is added to her background knowledge.

The irrelevance of a conditional credences to Jeffrey experience could be defined in a similar way. Suppose that an agent undergoes Jeffrey experience regarding $\mathbb{E}$. Let us think about the following equations:

(3.2) $\quad p(X|YE_1) = \cdots = p(X|YE_n) = p(X|Y)$.

These equations have the same spirit as (3.1). That is, (3.2) says that the agent's conditional credence in X given Y remains the same no matter which member of $\mathbb{E}$ is added to her background knowledge. Then, we can define the irrelevance of an agent's conditional credence to Jeffrey experience as follows:[8]

> **Definition 2:** Jeffrey experience regarding $\mathbb{E}$ *is irrelevant to an agent's conditional credence in X given Y* if and only if for all E_i, $p(XE_i|Y) = p(X|Y)p(E_i|Y)$.

Note that Definitions 1 and 2 heavily depend on the probabilistic independence relations. Diaconis and Zabell (1982) suggests a similar formulation as a definition of the probabilis-

[8] It is straightforward that when $p(E_i|Y) > 0$ for any $E_i \in \mathbb{E}$, (3.2) is equivalent to

(*) $\quad p(XE_i|Y) = p(X|Y)p(E_i|Y)$ for all E_i.

Note that (*) is more general than (3.2). In particular, when $p(E_i|Y) = 0$ for some $E_i \in \mathbb{E}$, (3.2) must have some undefined terms but (*) does not.

tic independence between a partition $\mathbb{E}$ and a proposition X. It is also noteworthy that Definition 2 is a general version of Definition 1.

Finally, we should provide a definition that can handle Adams experience. In other words, we should consider a course of experience that directly changes some conditional credences and nothing else. For simplicity, let's suppose that an agent undergoes Adams experience regarding $\{E, \bar{E}\}$-given-F. Here it should be emphasized that the change of the conditional credences in E given F and in $\bar{E}$ given F amounts to the change of the ratio between the credences in EF and in F. To put it in another way, the ratio between $p(EF)$ and $p(F)$ determines $p(E|F)$ and $p(\bar{E}|F)$, and *vice versa*. Then, one may attempt to define the irrelevance concerning Adams experience by means of the following equations:

(3.3) $\quad p(X|YEF) = p(X|YF) = p(X|Y)$.

At first sight, (3.1), (3.2) and (3.3) seem to share a common epistemic feature. That is, (3.3) says, like (3.1) and (3.2), that the agent's conditional credence in X given Y remains the same no matter which of EF and F is added to her background knowledge. Indeed, (3.3) is strong enough to be a sufficient condition for the conditional credence in X given Y to be irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F. The credences in EF and in F jointly determine the conditional credences in E given F (and in $\bar{E}$ given F). Thus, it sounds natural to state that the experience irrelevant to the credences in EF and in F is also irrelevant to the conditional credence in E given F and in $\bar{E}$ given F.

However, (3.3) is too strong to be a necessary condition for the irrelevance in question. Note that the conditional credences in E given F *does not* determine the credences in EF and in F. That is, $p(E|F)$ can determine only the ratio between the ratio between $p(EF)$ and $p(F)$, but not the values of $p(EF)$ and $p(F)$ themselves. So, it is implausible that the experience irrelevant to the conditional credences in E given F (and in $\bar{E}$ given F) is also irrelevant to the conditional credence in EF and in F.

To see this more clearly, let us compare two kinds of experience: Jeffrey experience regarding $\{EF, \bar{E}F, \bar{F}\}$ and Adams experience regarding $\{E, \bar{E}\}$-given-F. Note that (3.3) is equivalent to

(3.4) $\quad p(XEF|Y) = p(X|Y)p(EF|Y)$, $p(X\bar{E}F|Y) = p(X|Y)p(\bar{E}F|Y)$, and

$$p(X\bar{F}|Y) = p(X|Y)p(\bar{F}|Y).^{9}$$

According to Definition 2, then, the conditional credence in X given Y satisfying (3.3) is irrelevant to Jeffrey experience regarding $\{EF, \bar{E}F, \bar{F}\}$. So, if the irrelevance concerning Adams experience regarding $\{E, \bar{E}\}$-given-F is defined using only (3.3), then the conditional credences irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F should be regarded as being irrelevant to Jeffrey experience regarding $\{EF, \bar{E}F, \bar{F}\}$. However, there are some (conditional) credences that are irrelevant to the former but not the latter. To see this, let's take into account the relationship between the (conditional) credence in F (given $\mathbf{T}$) and Adams experience regarding $\{E, \bar{E}\}$-given-F. Here it should be emphasized that the conditional credence in E given F carries no information about the credence in F. In other words, for any real numbers $a \in [0, 1]$ and $b \in (0, 1]$, the proposition that the conditional credence in E given F equals a is compatible with the proposition that the credence in F equals b.[10] Thus, it sounds intuitive to say that the credence in F is irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F. Then, how about the relationship between the credence in F and Jeffrey experience regarding $\{EF, \bar{E}F, \bar{F}\}$? It is intuitively very clear that the credence in F is intimately related with Jeffrey experience that directly changes the credences in $\bar{F}$. Moreover, Definition 2 also says that they are not irrelevant to each other.[11] As a result, the (conditional) credence in F (given $\mathbf{T}$) is irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F, but not Jeffrey experience regarding $\{EF, \bar{E}F, \bar{F}\}$. In conclusion, (3.3) is a sufficient but not necessary condition for the conditional credence in X given Y to be irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F.

Then, how can we define the irrelevance at issue? Think about the following equations:

(3.5) $\quad p(E|XYF) = p(E|YF) = p(E|F).$

[9] Here it is assumed that $p(EF|Y) > 0$, $p(\bar{E}F|Y) > 0$ and $p(\bar{F}|Y) > 0$. Note that (3.3) is equivalent to

(3.3*) $\quad p(X|YEF) = p(X|Y\bar{E}F) = p(X|Y\bar{F}) = p(X|Y).$

Then, it follows from the definition of conditional credences that (3.3*) is equivalent to (3.4).

[10] According the standard probability theory, when the credence in F equals zero, the conditional credence in E given F is not well-defined. Thus, the standard theory says that the proposition that the credence in F equals zero is not compatible with the proposition that the conditional credence in E given F equals a particular real number. In order to rule out this possibility, the domain of the valuable b should be $(0, 1]$, not $[0, 1]$.

[11] Setting X and Y in Definition 2 to be F and $\mathbf{T}$ respectively, it holds that $p(FEF|\mathbf{T}) \neq p(F|\mathbf{T})p(EF|\mathbf{T})$, $p(F\bar{E}F|\mathbf{T}) \neq p(F|\mathbf{T})p(\bar{E}F|\mathbf{T})$ and $p(F\bar{F}|\mathbf{T}) \neq p(F|\mathbf{T})p(\bar{F}|\mathbf{T})$, when $p(F) \neq 1$.

First, it should be noted that (3.3) is not equivalent to (3.5).[12] But I think that (3.5) is another sufficient condition for the irrelevance at issue. Suppose that for any X, Y, E and F, (3.3) is a sufficient condition for the conditional credence in X given Y to be irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F. Then, it follows from (3.5) that the conditional credence in E given F is irrelevant to Adams experience regarding $\{X, \bar{X}\}$-given-Y. I assume here that the conditional credence in E given F is irrelevant to Adams experience regarding $\{X, \bar{X}\}$-given-Y if and only if the conditional credence in X given Y is irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F. Keeping it in mind that irrelevance relations are generally regarded as symmetric, we could find this assumption plausible. Then, (3.5) implies that the conditional credence in X given Y is irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F. To sum up, if (3.3) is a sufficient condition for the conditional credence in X given Y to be irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F, then so should (3.5). Moreover, setting X and Y in (3.5) to be F and $\mathbf{T}$ respectively, we can immediately ascertain that (3.5) holds. Thus, if (3.5) is another sufficient condition for a conditional credence in X given Y to be irrelevant to Adams experience regarding $\{E, \bar{E}\}$-given-F, then the (conditional) credence in F given $\mathbf{T}$ can be classified as irrelevant to the experience in question, which is in keeping with our intuition. This result cannot be obtained if the irrelevance regarding Adams experience is defined using only (3.3).

In light of the consideration, I propose to define the irrelevance in question as follows:

> **Definition 3:** Adams experience regarding $\{E, \bar{E}\}$-given-F *is irrelevant to her conditional credence in* X *given* Y if and only if
>
> (3a) $p(XF|Y) = p(X|Y)p(F|Y)$ and $p(XEF|Y) = p(X|Y)p(EF|Y)$; or
>
> (3b) $p(EY|F) = p(E|F)p(Y|F)$ and $p(EXY|F) = p(E|F)p(XY|F)$.

When $p(EF|Y) \neq 0$, (3a) is equivalent to (3.3). And when $p(EF|Y) = 0$, (3.3) has some undefined terms but (3a) does not. Unlike (3.3), thus, (3a) helps us to handle some cases

[12] Suppose that $p(XYEF) = p(XY\bar{E}F) = p(\bar{X}YEF) = p(\bar{X}YE\bar{F}) = p(\bar{X}Y\bar{E}F) = p(\bar{X}Y\bar{E}\bar{F}) = 0.1$ and $p(XYE\bar{F}) = p(\bar{Y}EF) = p(\bar{Y}E\bar{F}) = p(\bar{Y}\bar{E}F) = p(\bar{Y}\bar{E}\bar{F}) = 0.08$. Then, we have that $p(E|XYF) = p(E|YF) = p(E|F) = 0.5$ while $p(X|EFY) = p(X|FY) = 0.5 \neq (7/17) = p(X|Y)$.

in which $p(EF|Y) = 0$. For this reason, I prefer (3a) to (3.3). A similar consideration goes with (3.5) and (3b). Now we can generalize Definition 3 as follows:

> **Definition 4:** Adams experience regarding $\mathbb{E}$-given-F *is irrelevant to her conditional credence in* X *given* Y if and only if
>
> (4a) for all E_i, $p(XF|Y) = p(X|Y)p(F|Y)$ and $p(XE_iF|Y) = p(X|Y)p(E_iF|Y)$; or
>
> (4b) for all E_i, $p(E_iY|F) = p(E_i|F)p(Y|F)$ and $p(E_iXY|F) = p(E_i|F)p(XY|F)$.

This definition is a general version of Definition 2. In particular, setting F in Definition 4 to be $\mathbf{T}$, we have the same definition as Definition 2.[13]

With the help of the above definitions, we can re-formulate more explicitly Credal Conservatism as follows:

> **$\mathbf{CC_S}$**: When an agent undergoes simple experience regarding E,
>
> (S3) for all X and Y, $p(X|Y) = q(X|Y)$ if $p(XE|Y) = p(X|Y)p(E|Y)$.
>
> **$\mathbf{CC_J}$**: When an agent undergoes Jeffrey experience regarding $\mathbb{E}$,
>
> (J3) for all X and Y, $p(X|Y) = q(X|Y)$ if for all E_i, $p(XE_i|Y) = p(X|Y)p(E_i|Y)$.
>
> **$\mathbf{CC_A}$**: When an agent undergoes Adams experience regarding $\mathbb{E}$-*given*-F,
>
> (A3) for all X and Y, $p(X|Y) = q(X|Y)$ if
>
> (a) for all E_i, $p(XF|Y) = p(X|Y)p(F|Y)$ and $p(XE_iF|Y) = p(X|Y)p(E_iF|Y)$; or
>
> (b) for all E_i, $p(E_iY|F) = p(E_i|F)p(Y|F)$ and $p(E_iXY|F) = p(E_i|F)p(XY|F)$.

[13] To see this, let's set F in Definition 4 to be $\mathbf{T}$. Then, we have that:

> **Definition 4*:** Adams experience regarding $\mathbb{E}$-given-$\mathbf{T}$ *is irrelevant to her conditional credence in* X *given* Y if and only if
>
> (4a*) for all E_i, $p(XE_i|Y) = p(X|Y)p(E_i|Y)$; or
>
> (4b*) for all E_i, $p(E_iY) = p(E_i)p(Y)$ and $p(E_iXY) = p(E_i)p(XY)$.

It can be easily shown that (4b*) implies (4a*). Thus, the disjunction of (4a*) and (4b*) is equivalent to (4a*). So, we have that:

> **Definition 4**:** Adams experience regarding $\mathbb{E}$-given-$\mathbf{T}$ *is irrelevant to her conditional credence in* X *given* Y if and only if for all E_i, $p(XE_i|Y) = p(X|Y)p(E_i|Y)$.

This is the same as Definition 2.

CC_S, CC_J and CC_A correspond to courses of experience that trigger SC, JC and AC belief updating, respectively. And (S3), (J3) and (A3) represent explicitly the irrelevance of a conditional credence to the experience. Note that these formulations are proposed with the help of probabilistic independence (and its variant) that is mathematically well-founded. Before finishing this section, I would like to emphasize the following two Propositions:[14]

Proposition 3: When an agent undergoes Simple experience regarding E, (J3) and (A3) each are equivalent to (S3).

Proposition 4: When an agent undergoes Jeffrey experience regarding $\mathbb{E}$, (A3) is equivalent to (J3).

With the help of these Propositions, we can say that CC_S is a special case of CC_J and CC_A, and that CC_J is a special case of CC_A.

4 The Equivalence of Conditionalization with Credal Conservatism

So far, I have defined, in a probabilistic way, the irrelevance of conditional credences to experience. And these definitions have helped us give flesh to CC, i.e. Credal Conservatism. Now, we are ready to derive our main result: CC is equivalent to Conditionalization. Exactly speaking, it will be proved in this section that CC_S, CC_J and CC_A are equivalent to SC, JC and AC, respectively. For this purpose, we will prove first that:

Proposition 5: AC is equivalent to CC_A.

As was already mentioned, AC is a generalized version of SC and JC, and CC_A is a generalized version of CC_S and CC_J. Thus, the proof of the equivalence of AC with CC_A would be of help in proving the equivalence of SC with CC_S, and of JC with CC_J.

To prove Proposition 5, it is sufficient to show that:

[14] Recall that Definition 4 is a general version of Definition 2, and that Definition 2 is a general version of Definition 1. Then, we can ascertain Propositions 3 and 4 with no difficulty.

Proposition 6 When an agent undergoes Adams experience regarding $\mathbb{E}$-*given-*F, (A1) is equivalent to (A3).

Assume that an agent undergoes Adams experience regarding $\mathbb{E}$-given-F. First, let's prove that (A3) implies (A1). As mentioned in Section 2, it is well-known that (A1) is equivalent to (A2). For the present purpose, thus, I will show that (A3) implies (A2). Setting X and Y in (A3) to be F and a tautology $\mathbf{T}$ respectively, (A3) immediately implies that $p(F) = q(F)$. Similarly, setting Y in (A3) to be $\bar{F}$, we can derive from (A3) that for all X, $p(X|\bar{F}) = q(X|\bar{F})$. Lastly, let's set Y in (A3) to be E_jF. (Here, E_j is an arbitrary member of $\mathbb{E}$.) Then (A3) implies that for all X, $p(X|E_jF) = q(X|E_jF)$. (See Appendix 1.) To sum up, (A3) implies that $p(F) = q(F)$; for all X, $p(X|\bar{F}) = q(X|\bar{F})$; and for all X and $E_i(\in \mathbb{E})$, $p(X|E_iF) = q(X|E_iF)$. That is, (A3) implies (A2) that is equivalent to (A1).

Now, let's show that (A1) implies (A3). According to (A1), it holds for two arbitrary propositions X and Y that:

$$(4.1) \quad q(X|Y) = \frac{q(XY)}{q(Y)} = \frac{\Sigma_i q(E_i|F)p(XY|E_iF)p(F)+p(XY\bar{F})}{\Sigma_i q(E_i|F)p(Y|E_iF)p(F)+p(Y\bar{F})}.^{15}$$

Assume that the antecedent of (A3) holds for the X and Y in (4.1). That is, assume that:

(4.2) for all E_i, $p(X\bar{F}|Y) = p(X|Y)p(\bar{F}|Y)$ and $p(XE_iF|Y) = p(X|Y)p(E_iF|Y)$, or

(4.3) for all E_i, $p(E_iY|F) = p(E_i|F)p(Y|F)$ and $p(E_iXY|F) = p(E_i|F)p(XY|F)$.

Here I should point out that when (4.1) holds, (4.2) and (4.3) each imply that $p(X|Y) = q(X|Y)$. (The relevant proofs are given in Appendix 2 .) Note that $p(X|Y) = q(X|Y)$ is the consequent of (A3). Thus, we can say that (A1) implies (A3). As a result, we have that (A1) is equivalent to (A3), and thus that AC is equivalent to $\mathrm{CC_A}$. *Q.E.D.*

Note that this result immediately implies the following proposition:[16]

Proposition 7: SC and JC are equivalent to $\mathrm{CC_S}$ and $\mathrm{CC_J}$, respectively.

[15] It is assumed here that $0 < p(Y) < 1$ and $0 < p(E_iF) < 1$ for all $E_i \in \mathbb{E}$. These assumptions and AC jointly imply that $0 < q(Y) < 1$, and so that $q(X|Y)$ is well-defined.

[16] Suppose that an agent undergoes Simple experience regarding E. Then, Propositions 1 and 3 imply that (A1) is equivalent to (S1), and that (A3) is equivalent to (S3). On the other hand, according to Proposition 6, (A1) is equivalent to (A3). Hence, we have that when an agent undergoes Simple experience regarding E, (S1) is equivalent to (S3). In other words, we obtain that SC is equivalent to $\mathrm{CC_S}$. In a similar way, it can be proved that JC is equivalent to $\mathrm{CC_J}$.

This result is interesting. Consider JC, for example. All JC requires is that the new credence in X should be obtained from the old conditional credences in X given E_i by taking a weighted average, with the weights being the new credences in E_i. However, this requirement itself is not helpful for us to understand clearly the epistemological traits of JC. Thus, some equivalent formulations of SC and JC have hitherto been suggested. (See Jeffrey (1992, pp.118-126; 2004, pp.51-55).) For example, (J1) is equivalent to:

> For all X, Y and E_i, $P(X)/P(Y) = Q(X)/Q(Y)$ if each of X and Y implies E_i.
>
> For all X and E_i, $Q(E_i)/P(E_i) = Q(X)/P(X)$ if X implies E_i.

Similar to (J2), which is often called Rigidity, these formulations tell what should remain the same when our credences are updated in accordance with JC. However, the equivalent formulations fail to show clearly the epistemological plausibility of JC. Even Jeffrey (1992, p.118), it seems, thought that Rigidity and the above equivalent formulations are somewhat unclear. Why should the ratios in question remain the same after we undergo some relevant experience? It is not easy to give a satisfactory response to this question. Unlike these equivalent formulations, however, $\mathrm{CC_J}$ displays clearly an epistemic norm—namely, epistemic conservatism. It is hard to deny, thus, that $\mathrm{CC_J}$ shows more transparently some epistemological traits of JC than those equivalent formulations. The like may be true of SC and AC. Interestingly, very little attention has been paid to the equivalence between CC and Conditionalization as far as I know. Thus, Propositions 5 and 7 could be regarded as a substantial contribution to Bayesian epistemology.[17]

[17] Here I should mention that Dietrich, List and Bradley (unpublished) recently provide some arguments that have some similar features to mine. That is, they define 'Conservativeness' in a formal way, and examine how it is related with Bayesian updating rules. However, there are some substantial differences between their work and mine. First, my Credal Conservatism is formulated by means of *one credence (or probability) function* that an agent had before undergoing some experience while their Conservativeness is formulated by means of *some sets of probability functions* that represent an agent's belief states and some experience that she undergoes. Second, my main result is that Credal Conservatism *is equivalent to* Conditionalization while Dietrich, List and Bradley show just that (when their 'Responsiveness' is assumed,) Conservativeness *implies* Conditionalization.

5 Concluding Remarks

Roughly, I have shown in this paper that our credences are updated in accordance with Credal Conservatism if and only if our credences are updated using the Bayesian updating rule—i.e. Conditionalization. If Credal Conservatism is a plausible epistemic norm, then this result would be regarded as a justification of SC, JC and AC. I don't rule out the possibility, however, that one provides powerful arguments against Credal Conservatism. Nonetheless, it cannot be denied that, unlike existing equivalent formulations of Conditionalization, Credal Conservatism shows clearly an important epistemological trait of the Bayesian updating rule.

Appendix 1

First, let's set X and Y in (A3) to be F and a tautology $\mathbf{T}$ respectively. Then, (A3) immediately implies that:

(I) $p(F) = q(F)$ if

(I.1) for all E_i, $p(F|\mathbf{T}) = p(F|\mathbf{T})p(F|\mathbf{T})$ and $p(FE_iF|\mathbf{T}) = p(F|\mathbf{T})p(E_iF|\mathbf{T})$;
or

(I.2) for all E_i, $p(E_i\mathbf{T}|F) = p(E_i|F)p(\mathbf{T}|F)$ and $p(E_iF\mathbf{T}|F) = p(E_i|F)p(F\mathbf{T}|F)$.

Note that (I.2) must be true. Hence, the antecedent of (I) holds, and so we have that $p(F) = q(F)$. Second, let's set Y in (A3) to be $\bar{F}$. Then, it follows from (A3) that:

(II) For all X, $p(X|\bar{F}) = q(X|\bar{F})$ if

(II.1) for all E_i, $p(XF|\bar{F}) = p(X|\bar{F})p(F|\bar{F})$ and $p(XE_iF|\bar{F}) = p(X|\bar{F})p(E_iF|\bar{F})$;
or

(II.2) for all E_i, $p(E_i\bar{F}|F) = p(E_i|F)p(\bar{F}|F)$ and $p(E_iX\bar{F}|F) = p(E_i|F)p(X\bar{F}|F)$.

It is obvious that both (II.1) and (II.2) must be true. This is because for any Z, $p(ZF|\bar{F}) = p(Z\bar{F}|F) = 0$. Hence, we have that for all X, $p(X|\bar{F}) = q(X|\bar{F})$. Lastly, let's set Y in (A3) to be E_jF. (Here, E_j is arbitrary member of $\mathbb{E}$.) Then (A3) implies that:

(III) For all X, $p(X|E_jF) = q(X|E_jF)$ if

(III.1) for all E_i, $p(XF|E_jF) = p(X|E_jF)p(F|E_jF)$ and $p(XE_iF|E_jF) = p(X|E_jF)p(E_iF|E_jF)$; or

(III.2) for all E_i, $p(E_iE_jF|F) = p(E_i|F)p(E_jF|F)$ and $p(E_iXE_jF|F) = p(E_i|F)p(XE_jF|F)$.

Clearly, (III.1) must holds for all X. Above all, note that it must be true that $p(XF|E_jF) = p(X|E_jF)p(F|E_jF)$. This is because $p(XF|E_jF) = p(X|E_jF)$ and $p(F|E_jF) = 1$. How about the proposition that for all E_i, $p(XE_iF|E_jF) = p(X|E_jF)p(E_iF|E_jF)$? First, consider the cases in which $i \neq j$. Then, we have that $p(XE_iF|E_jF) = 0 = p(X|E_jF)p(E_iF|E_jF)$. This is because the members of $\mathbb{E}$ are mutually exclusive and so $p(XE_iF|E_jF) = p(E_iF|E_jF) = 0$ when $i \neq j$. Second, consider the cases in which $i = j$. In these cases, it holds that:

$p(XE_jF|E_iF) = p(XE_iF|E_iF) = p(X|E_iF)$ and
$p(X|E_jF)p(E_iF|E_jF) = p(X|E_iF)p(E_iF|E_iF) = p(X|E_iF)$.

Thus, $p(XE_iF|E_jF) = p(X|E_iF)p(E_iF|E_jF)$ when $i = j$. As a result, we obtain that for all E_i, $p(XE_iF|E_jF) = p(X|E_jF)p(E_iF|E_jF)$. Thus, the second conjunct of (III.1) is also true. Thus, we have that (III.1) holds for all X, and so it follows from (III) that for all X, $p(X|E_jF) = q(X|E_jF)$.

Appendix 2

Let's prove that when (4.1) holds, (4.2) and (4.3) each imply that $p(X|Y) = q(X|Y)$. That is, I will show that:

I. (4.1) and (4.2) jointly imply that $p(X|Y) = q(X|Y)$, and

II. (4.1) and (4.3) jointly imply that $p(X|Y) = q(X|Y)$.

First, let me show that (4.1) and (4.2) jointly imply that $p(X|Y) = q(X|Y)$. Suppose that (4.2) holds. That is, let's assume that:

(4.2) for all E_i, $p(X\bar{F}|Y) = p(X|Y)p(\bar{F}|Y)$ and $p(XE_iF|Y) = p(X|Y)p(E_iF|Y)$.

Then, we have that:

(4.2a) for all E_i, $p(XY\bar{F}) = p(X|Y)p(Y\bar{F})$ and $P(XY|E_iF) = p(X|Y)p(Y|E_iF)$.

With the help of the probability calculus and the first conjunct of (4.2), we have that:

$$p(XY\bar{F}) = p(X|Y)p(\bar{F}|Y)p(Y) = p(X|Y)p(Y\bar{F}).$$

That is, we obtain that $p(XY\bar{F}) = p(X|Y)p(Y\bar{F})$. Now, let's consider an arbitrary E_i. The probability calculus implies that:

$$p(XY|E_iF) = \tfrac{p(XE_iF|Y)p(Y)}{p(E_iF)}.$$

Then, Bayes theorem and the second conjunct of (4.2) imply that:

$$p(XY|E_iF) = \tfrac{p(XE_iF|Y)p(Y)}{p(E_iF)} = \tfrac{p(X|Y)p(E_iF|Y)p(Y)}{p(E_iF)} = p(X|Y)p(Y|E_iF).$$

That is, we obtain that for all E_i, $P(XY|E_iF) = P(Y|E_iF)P(X|Y)$. As a result, we have (4.2a) with the help of (4.2) and the probability calculus. Lastly, we should note that (4.1) and (4.2a) imply that:

$$\begin{aligned} q(X|Y) &= \frac{\Sigma_i q(E_i|F)p(XY|E_iF)p(F) + p(XY\bar{F})}{\Sigma_i q(E_i|F)p(Y|E_iF)p(F) + p(Y\bar{F})} \\ &= \frac{\Sigma_i q(E_i|F)p(X|Y)p(Y|E_iF)p(F) + p(X|Y)p(Y\bar{F})}{\Sigma_i q(E_i|F)p(Y|E_iF)p(F) + p(Y\bar{F})} = p(X|Y). \end{aligned}$$

Therefore, it can be concluded that (4.1) and (4.2) jointly imply that $p(X|Y) = q(X|Y)$. Second, let's prove that (4.1) and (4.3) jointly imply that $p(X|Y) = q(X|Y)$. Suppose that (4.3) holds. That is, let's assume that:

(4.3) for all E_i, $p(E_iY|F) = p(E_i|F)p(Y|F)$ and $p(E_iXY|F) = p(E_i|F)p(XY|F)$.

Note that (4.3) is equivalent to

(4.3a) for all E_i, $p(Y|E_iF) = p(Y|F)$ and $p(XY|E_iF) = p(XY|F)$.

Note also that $\Sigma_i q(E_i|F) = 1$. Then, we obtain from (4.1) and (4.3a) that:

$$\begin{aligned} q(X|Y) &= \frac{\Sigma_i q(E_i|F)p(XY|E_iF)p(F) + p(XY\bar{F})}{\Sigma_i q(E_i|F)p(Y|E_iF)p(F) + p(Y\bar{F})} \\ &= \frac{p(XY|F)p(F)\Sigma_i q(E_i|F) + p(XY\bar{F})}{p(Y|F)p(F)\Sigma_i q(E_i|F) + p(Y\bar{F})} = \frac{p(XY)}{p(Y)} = p(X|Y). \end{aligned}$$

Therefore, it can be concluded that (4.1) and (4.3) jointly imply that $p(X|Y) = q(X|Y)$.

References

Bradley, B. (2005) "Radical Probabilism and Bayesian Conditioning", *Philosophy of Science* 72: 342–364.

Christensen, D. (1994) "Conservatism in epistemology", *Noûs* 28: 69–89.

Diaconis, P. and Zabell, S.L. (1982) "Updating subjective probability", *Journal of the American Statistical Association* 77: 822–830.

Dietrich, F. List, C. Bradley, R. (unpublished), "A Joint Characterization of Belief Revision Rules."

Harper, W.L. and Kyburg, H.E. (1968) "The Jones case", *The British Journal for the Philosophy of Science* 19: 247–251.

Howson and Urbach (1993) *Scientific Reasoning: The Bayesian Approach*, 2nd ed. Open Court.

Jeffrey, R. (1992) *Probability and the Art of Judgment.* Cambridge University Press.

Jeffrey (2004) *Subjective probability: The real thing.* Cambridge University Press.

Levi, I. (1969) "If Jones only knew more!", *The British Journal for the Philosophy of Science* 20: 153–159.

Quine, W.V.O. (1951) "Two Dogmas of Empiricism," in *From a Logical Point of View*, 2nd ed. New York: Harper & Row, 1961.

Talbott, W. (2008) "Bayesian epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2008), http://plato. stanford. edu/archives/fall2008/entries/epistemology-bayesian.

van Fraassen, B. C. (1989). *Laws and symmetry.* Oxford: Clarendon Press.

Wagner (2003) "Commuting probability revisions: The uniformity rule". *Erkenntnis* 59: 349–364.

Quantum Mechanics for Event Ontologists

Thomas Pashby*

What is quantum mechanics about? That is, what is the intended domain of an interpretation of the theory? In the long history of attempts to interpret quantum theory a wide variety of answers have been given to this question, including: observations, experiments, wavefunctions, the universe, point particles, information. In this paper I explore a particular view of quantum mechanics which maintains that it is a theory of *events*. On this event ontology view, the probabilities supplied by the quantum state are probabilities for the occurrence of events, and the observables of the theory are to be interpreted accordingly. In contrast, the standard (Dirac-von Neumann) view maintains that observables correspond to physical quantities which, when measured, come to have definite values—values that represent possessed properties of the system. In this vein, Wightman (1962) interprets position experiments in terms of localization: on measurement, a particle comes to be localized within a particular region of space; it has the property of existing *here* rather then somewhere else.

The event view has had some recent interest from notable theoretical physicists such as Carlo Rovelli and Rudolf Haag, who express in different ways the core idea of this view. Rovelli (2005) contrasts the "'wave function ontology," which takes the state "as the 'real' entity which fully represents the actual state of affairs of the world," with his proposal for an "ontology of quantum events."

*Any questions or comments may be addressed to: tom.pashby@gmail.com.

> A better alternative is to take the observed values . . . as the actual elements of reality, and view [the state] as a mere bookkeeping device, determined by the actual values . . . that happened in [the] past. From this perspective, the real events of the world are the 'realizations' (the 'coming to reality', the 'actualization') of the values . . . in the course of the interaction between physical systems. These quantum events have an intrinsically discrete (quantized) granular structure. (p. 115)

The key idea is that the changing quantum state given by the dynamics of the theory does not describe the changing properties of some physical object.[1] Rather, the quantum state describes the probabilities for events to occur; events that often arise as the result of interactions between systems.

In turn, here is Haag's (2013) recent critique of the conventional view:

> What do we detect? The presence of a particle? Or the occurrence of a microscopic event? We must decide for the latter.. . . [T]he standard use of the term "observable" does not really correspond to the needs of collision theory in particle physics. We do not measure a "property of a microscopic system", characterized by a spectral projector of a self-adjoint operator. Rather we are interested in the detection of a microscopic event. The first task is to characterize the mutually exclusive alternatives for such an event. (p. 1310)

So in practice, i.e. in the context of a particle detection experiment, the theory concerns—is about—microscopic events, such as the ionization of a molecule by a cosmic ray. The detector is

[1] In approvingly citing Rovelli's commitment to an ontology of events I do not mean to endorse his accompanying interpretation of quantum theory, Relational Quantum Mechanics, about which there is much to criticize. I shall not do so here, however.

expressly designed to amplify these micro-events such that they reliably lead to a macroscopic record of detection, by which we mean detection at a place, at a *time*.

My main concern in this paper is the task of re-interpreting the spectral projectors of position in terms of events, and constructing an appropriate probability space for these events. That is, my focus will be on the re-interpretation of Wightman localization in terms of the occurrence of localized events within extended spatio-temporal regions. I claim that, interpreted in terms of events, there is a crucial further question concerning localization to which quantum mechanics must supply an answer: *When* does an event occur? Providing a satisfactory answer to this question, I contend, gives an informative account of Haag's 'Principle of Random Realization' and thus avoid Rovelli's paradox that "the statement that a certain specific outcome 'has happened' can be true and not-true at the same time" (p. 115).

The paper is laid out as follows. In Section 1, I provide an account of Wightman localization. In Section 2, I present a result that makes trouble for the conventional account of localization in terms of the possession of a property. In Section 3, I explore the idea that a localization experiment can be interpreted as a conditional probability for an event to occur in some region of space, given that it occurs at some time. I give an expression for these conditional probabilities in terms of Lüders Rule by forming a temporally extended space of histories of a system. In conclusion, I propose a philosophical interpretation of these probabilities as Lewisian objective chances, conditioned on the future occurrence of an event in the possible worlds to which the chances apply, but maintain that there is no need to adopt a corresponding Everettian 'no collapse' interpretation of quantum mechanics to accommodate them.

1. Localization as a Property The position of a quantum particle is given by the position observable, Q. But, considered as an operator, Q returns the *expectation value* of position. That

is, a system in state ψ (a unit vector in $\mathcal{H}$, a Hilbert space) has an expectation value for position $\langle\psi|Q\psi\rangle$. This does not refer to the position of a particular system, or a particular run of a position experiment, but rather a certain characteristic of the likely distribution for an ensemble of such systems. In the conventional interpretation, probabilistic statements about individual systems take the form: 'The value of position lies in the interval $[a, b]$.' This statement is associated with a spectral projector of Q, $P^Q_{[a,b]}$, through the Spectral Theorem, and the probability that, on measurement, the statement is found to be true is $\langle\psi|P_{[a,b]}\psi\rangle$. If such a statement is true, i.e. if $\langle\psi|P_{[a,b]}\psi\rangle = 1$, then this has been interpreted as saying that the system is located within the spatial region corresponding to $[a, b]$.

Wightman (1962) showed that directly associating a projector P_Δ with a spatial region Δ, and demanding of these projectors that they be appropriately related under symmetry transformations of the underlying space-time, suffices to determine uniquely the position operator Q. In his "notion of localizability in a region" Δ, these projectors are "supposed to describe a *property* of the system, the property of being localized in Δ" (p. 847, original emphasis). A localization experiment corresponds to a series of experimental questions[2] which ask "is the particle located in Δ?" One of the crucial features of such questions is that they must be asked at a particular instant of time so that quantum mechanics may provide answers via the Born Rule.

That is, a projector P_Δ associates with a state of the system ψ a vector $P_\Delta\psi \in \mathcal{H}$. According to the Born Rule, the probability of finding a system in state ψ to be located in the region Δ is given by the inner product $\langle\psi|P_\Delta\psi\rangle$. Thus if the probability is one, it must be the case that $P_\Delta\psi = \psi$. This is the eigenstate-eigenvalue link: if the system is located in Δ then the state ψ is an eigenstate of the projection P_Δ, and if the state is such an eigenstate of P_Δ then it is located in

[2]This term is due to Mackey who provided the more general notion of a System of Imprimitivity, of which Wightman's system of localization is an example.

Δ with probability one. Introducing time through the Schrödinger picture, we associate a time indexed state $\psi_t \in \mathcal{H}$ with a time t through the unitary group $U_t = e^{iHt}$ by setting $\psi_t = U_t\psi$, where H is the Hamiltonian operator and ψ is the state at $t = 0$.

Operating on these time-indexed states, the projector returns a vector $P_\Delta\psi_t \in \mathcal{H}$, which (properly normalized) is a state located in Δ at the time t. The probability of localization in Δ at a time t is, therefore, $\langle\psi_t|P_\Delta\psi_t\rangle$. However, in general a state such as this will fail to be localized in Δ at any other time. In fact, we can say something quite definitive about the character of the times at which a state can be localized in this way. To do so requires the Heisenberg picture, which is reached by considering the time evolution of the observables rather than the state. That is, the Heisenberg projector corresponding to localization in Δ at time t is $P_\Delta(t) = U_{-t}P_\Delta U_t$, and the Heisenberg state of the same system is $\psi = \psi_0$, for all t.

Thus the probabilities supplied by the Heisenberg picture are numerically identical to those of the Schrödinger picture since $\langle\psi|P_\Delta(t)\psi\rangle = \langle\psi|U_{-t}P_\Delta U_t\psi\rangle = \langle\psi_t|P_\Delta\psi_t\rangle$. Conceptually, however, it is now more straightforward to associate properties of the system possessed at distinct times with the state of the system described by a single vector $\psi \in \mathcal{H}$. A system localized in region Δ_1 at t_1 and Δ_2 at t_2 must be associated with a state ψ such that $P_{\Delta_1}(t_1)\psi = P_{\Delta_2}(t_2)\psi = \psi$.[3]

2. A Problematic Result I will now show that the times at which a system is localized in this way are severely limited, so long as the Hamiltonian of the system obeys a common requirement known as the Spectrum Condition, which requires the system to have lowest value of energy below which it cannot drop. If this condition holds, as it does for all physically reasonable

[3]This amounts to the assumption that $P_{\Delta_1}(t_1)P_{\Delta_2}(t_2)\psi = P_{\Delta_2}(t_2)P_{\Delta_1}(t_1)\psi$, which won't be true in general (Malament 1996). This difficulty is closely related to the problem I display below.

quantum systems, then there can be no time interval such that a system is localized in some region at every time in that interval, unless it is localized within that region at *all* times. This follows from the following proposition.

Proposition 1. *Let P be a projection operator associated with a property of the system, let $\psi \in \mathcal{H}$ be a vector of unit norm in a separable Hilbert space, and let $U_t = e^{iHt}$ be the one-parameter unitary group uniquely generated by H, a self-adjoint operator with semi-bounded spectrum. Let $P(\{t_k\})$ be the projection operator that corresponds to the possession of the property P at every time $t \in \{t_k\}$. That is, ψ is in the range of $P(\{t_k\})$ if $P(t)\psi = \psi$ for all $t \in \{t_k\}$. Let ψ be in the range of $P(\{t_k\})$ then* either*:*

1. *$\{t_k\}$ is a set with zero Lebesgue measure,* or
2. *ψ is in the range of $P(\mathbb{R})$, i.e. $P(\{t_k\}) = P(\mathbb{R})$.*

Therefore, there is no projection $P(I)$ that corresponds to the possession of a property at an open interval of instants $I \subset \mathbb{R}$ and at no other time.

Let us consider the implications of this result in the context of localization. Applied to the projector P_Δ, the result says that the states which possess the property of being localized in Δ at more than one time are severely limited. One way of interpreting the result would be to say that if a system is confined to a region of space for a time interval (i.e. a continuous set of instants) then it is confined to that region for all time. This resembles the Quantum Zeno (or Watchdog) Effect where continuous measurement of a projector confines the evolution of the system to a subspace of its state space (Misra & Sudarshan 1977).

If, on the other hand, we have reason to believe that the system is not localized in some region for all time, then the times at which the system is localized *anywhere* within that region are

severely limited. This is because the result applies to subregions as well. That is, if $\Sigma \supset \Delta$ is a larger region that includes Δ then $P_\Sigma(t)\psi \neq \psi$ implies that $P_\Delta(t)\psi \neq \psi$.[4] So the conclusion of the result restricts localization within a subregion of the region just as much as it does for the region itself, and the region under consideration could be as large as one likes: the Earth, the Solar System, or so on. Moreover, the condition that a state must satisfy to be localized within a region for a time interval is so severe that a non-zero *probability* that the system is localized anywhere outside of the region at *any* time t entails that the set of times at which it is localized within the region has measure zero.

The upshot of all this is that if we are to admit the mere possibility that the particle could be detected outside of the lab next week, then it cannot be localized within the lab today at more than a set of times with zero measure. Therefore, on the Wightman interpretation, the properties of systems (or at least the properties that we can hope to have empirical access to) are temporally *sparse*, in this sense. But what sort of persisting physical object fails to have spatial properties (in the regions we care about) at the vast majority of times? This is, I claim, a further indication that this interpretation of the state, as describing the changing properties of a physical thing, is a mistake. In its place, I propose an account of localization in terms of the occurrence of spatio-temporally located events rather than possessed properties.

3. Localization as Occurrence of an Event Picture a typical diffraction experiment which involves a source emitting a beam of particles, a diffraction grating through which the beam passes, and a luminescent screen. The source of quanta (electrons or photons, say) emits a single quantum particle at a time, at a frequency such that only a single particle is ever in the apparatus.

[4]Let $P_\Delta(t)\psi = \psi$ then, since $P_\Sigma(t)P_\Delta(t) = P_\Delta(t)$, we have $P_\Sigma(t)\psi = P_\Sigma(t)P_\Delta(t)\psi = P_\Delta(t)\psi = \psi$, in contradiction with the assumption made above.

Some time after a particle is emitted, a dot appears on the screen, and, repeating the experiment many times, the relative intensity of these discrete events comes to form a characteristic spatial interference pattern. Some things to note: first, the outcome of the experiment is an event, i.e. a definite occurrence situated in space and time; second, the time interval after which the dot appears will vary; finally, the screen is sensitive over the entire course of the experiment, and an individual experiment ends only when the particle is detected. Taken together, these observations suffice to show that the usual interpretation of localization as a property arising from an instantaneous measurement of position cannot be right.

Following Haag's suggestion (above), our first task is to characterize the mutually exclusive alternatives for our detection event. Clearly, a single event occurs in just one comparatively small region of the screen at one time. So our outcome space must allow for variation in space and time. Furthermore, since every run of the experiment ends with a detection, the elementary event (to which probability one is assigned) must correspond to a dot appearing somewhere on the screen at some time after emission. None of these characteristics mesh well with the idea that spatial localization is a property resulting from an instantaneous measurement of the corresponding projector. In particular, the elementary event for a localization experiment is always localization *at a time*.

As a first step, I propose that we interpret $P_\Delta(t)$ not as an experimental question asked at time t, but rather as a proposition about an event: the proposition that an event occurs in spatial region Δ at time t. David Lewis (1986) gave an account of an event as a property (or class) of spatio-temporal regions as follows, which is easily adapted to the present case.

> To any event there corresponds a property of regions: the property that belongs to all and only those spatio-temporal regions, of this or any other possible world, in which that event occurs. Such a property belongs to exactly one region of any world where

the event occurs . . . (p. 243)

Note first that Lewis distinguishes 'occurring in' a region from 'occurring within' a region. If an event occurs within a region Δ then, according to Lewis, it occurs within every super-region $\Sigma \supset \Delta$. This closely resembles the account of Wightman localization given in the previous sections since if $P_\Delta \psi = \psi$ then $P_\Sigma \psi = \psi$ for all $\Sigma \supset \Delta$. However, according to Lewis, an event occurring *in* a region Δ does not occur in any super-region, nor any subregion. The region in which an event occurs is, therefore, 'just the right size.' This can be achieved here by means of the following definition: *Given a state ψ, a localization event occurs in a region Δ if and only if $P_\Delta \psi = \psi$ and there is no subregion $\Omega \subset \Delta$ such that $P_\Omega \psi = \psi$.* This ensures that a localization event cannot occur in Δ and its super-region Σ, since if it occurs in Δ (and thus obeys the first condition) then the latter condition (required for occurrence in Σ) is not satisfied.

We can think of a certain class of states as giving the relevant space of possible worlds. We are interested in worlds in which a system prepared in (Heisenberg) state ψ results in a detection event, i.e. an event occurring within the screen some time after emission from the source. Let the detector be sensitive in region Δ over all times.[5] Then the possible outcomes of the experiment correspond to worlds where $P_\Delta(t)\psi = \psi$, i.e. worlds in which a detection event occurs at time t. The elementary event, to which we must assign probability one, is the occurrence of an event within Δ at some time $t \in \mathbb{R}$. This event (corresponding to a class of possible worlds) is associated with the state ψ_Δ for which $P_\Delta(t)\psi_\Delta = \psi_\Delta$ at every $t \in \mathbb{R}$.

Quantum mechanics gives conditional probabilities through Lüders' Rule, according to which

[5]Effectively, we assume here that emission occurs some time in the distant past, i.e. at $t = -\infty$.

the probability of outcome A given outcome B is[6]

$$\Pr(A|B) = \frac{\langle P_B\psi | P_A P_B \psi\rangle}{\langle \psi | P_B \psi\rangle},$$

where P_A and P_B are projectors representing outcomes A and B, respectively, such that $P_A \leq P_B$. But what *instantaneous* projector could correspond to our elementary event?

It is here that the limitations of the Heisenberg and Schrödinger pictures start to bite, since each considers only instantaneous projectors. But the relevant projectors here concern a continuous interval of time, e.g. the projector onto all states ψ_Δ such that $P_\Delta(t)\psi_\Delta = \psi_\Delta$ at every $t \in \mathbb{R}$.[7] The difficulty we face is that every vector $\psi \in \mathcal{H}$ uniquely corresponds to a history $\psi_t = U_t\psi$, but the histories we are interested in (corresponding to the occurrence of an event within some time interval) must be defined more generally.

To free ourselves from this restriction, let us consider instead vector valued functions of t, $\Psi(t) = \psi(t)$, with $\psi(t) \in \mathcal{H}$. Such a function represents an entire history of a system, i.e a possible world. We may thus define a history $\Psi_\Delta(t)$ corresponding to the desired elementary event by the function $\Psi_\Delta(t) = P_\Delta U_t \psi$. But these functions of t cannot lie in the instantaneous Hilbert space $\mathcal{H} = L^2[\mathbb{R}^3]$ of functions of (configuration) space. Instead, we must consider the *temporally extended* Hilbert space $\mathcal{H}_+ = L^2[\mathbb{R}^3] \times L^2[\mathbb{R}]$ of functions of space and time.[8] We can now define a projector P^+_Δ on $\mathcal{H}_+$ that operates on every instantaneous state, $P^+_\Delta \Psi(t) = P_\Delta \psi(t)$.

[6]Lüders' Rule is usually given in terms of the trace, but for simplicity's sake we will only consider pure states here.

[7]In one-dimension Δ is an interval $[a, b]$. The relevant subspace is the square integrable functions on the interval $[a, b]$, i.e. $\psi_\Delta \in L^2[a, b]$, states for which $P_\Delta(t)\psi_\Delta = \psi_\Delta$ (implying that the (sub-)domain of the Hamiltonian is closed in $L^2[a, b]$).

[8]See the Appendix for a rigorous definition of $\mathcal{H}_+$ as a continuous direct sum.

Moreover, we can define a projection operator $P^T([t_1, t_2])$ which has the effect of truncating an arbitrary history $\Psi(t)$ as follows:

$$P^T([t_1, t_2])\Psi(t) = \begin{cases} \psi(t) & \text{if } t_1 \leq t \leq t_2 \\ 0 & \text{otherwise.} \end{cases}$$

Armed with these projectors onto times, we may now associate the outcome space of our diffraction experiment with conditional probabilities through Lüders' Rule. Let $\psi \in \mathcal{H}$ be the Heisenberg state of the system in question and let $\Psi(t) = U_t\psi$. Then, making use of the inner product on $\mathcal{H}_+$ (see Appendix), Lüders' Rule returns probability one for the elementary event corresponding to detection within Δ at some t, as required,

$$\Pr(\Delta|\Delta) = \frac{\langle P^+_\Delta \Psi | P^+_\Delta P^+_\Delta \Psi\rangle_+}{\langle \Psi | P^+_\Delta \Psi\rangle_+} = \frac{\lim_{\tau\to\infty}\int_{-\tau}^{\tau}\langle\psi(t)|P_\Delta\psi(t)\rangle dt}{\lim_{\tau\to\infty}\int_{-\tau}^{\tau}\langle\psi(t)|P_\Delta\psi(t)\rangle dt} = 1.$$

But we may also obtain the conditional probability for detection during $I = [t_1, t_2]$ by means of the projector $P^T([t_1, t_2])$ defined above,

$$\Pr(I|\Delta) = \frac{\langle P^+_\Delta \Psi | P^T(I) P^+_\Delta \Psi\rangle_+}{\langle \Psi | P^+_\Delta \Psi\rangle_+} = \frac{\int_{t_1}^{t_2}\langle\psi(t)|P_\Delta\psi(t)\rangle dt}{\lim_{\tau\to\infty}\int_{-\tau}^{\tau}\langle\psi(t)|P_\Delta\psi(t)\rangle dt} < 1.$$

Thus we obtain the means to associate probabilities with the occurrence of events localized in time and space. From this perspective, the Heisenberg and Schrödinger pictures are rather limiting since a unit vector $\psi(t) \in \mathcal{H}$ can only be associated with an instantaneous elementary event, in which case having unit norm says that, with certainty, an event will occur at t. On the contrary, in defining the conditional probability $\Pr(I|\Delta)$ by an integral, we treat $\langle\psi(t)|P_\Delta\psi(t)\rangle$ as a probability *density* rather than a probability and so the probability of occurrence during any particular instant t is zero. This latter result seems to correctly reflect the probability that a detector sensitive for a mere instant would fire at that instant—real detectors are sensitive over time intervals, not collections of instants (with measure zero).

Before concluding it should be acknowledged that there are close links of my proposal to that of Brunetti & Fredenhagen (2002) for event time Positive Operator Valued Measures (POVMs), taken up by (e.g.) Hegerfeldt & Muga (2010). The crucial distinction, however, is that their interpretation of these POVMs does not serve to define a valid conditional probability (which assigns probability one to the occurrence of the elementary event described by the condition). The relation is as follows. If $I \mapsto F(I)$ is the POVM in $\mathcal{H}$ that they associate with an event occurring in Δ during I then my conditional probability (above) can be written as:

$$\Pr(I|\Delta) = \frac{\langle P^+_\Delta \Psi | P^T(I) P^+_\Delta \Psi \rangle_+}{\langle \Psi | P^+_\Delta \Psi \rangle_+} = \frac{\langle \psi | B(\mathbb{R})^{1/2} F(I) B(\mathbb{R})^{1/2} \psi \rangle}{\langle \psi | B(\mathbb{R}) \psi \rangle}$$

where $B(\mathbb{R})$ is the positive operator $B([-\infty, \infty]) = \int_{-\infty}^{\infty} P_\Delta(t) dt$, and $F(I) = B(\mathbb{R})^{-1/2} B(I) B(\mathbb{R})^{-1/2}$ with $B(I) = \int_{t_1}^{t_2} P_\Delta(t) dt$.

4. Conclusion The conditional probabilities obtained for these spatio-temporally localized events can be usefully thought of a Lewisian chances (Lewis 1981): the condition serves to pick out those possible worlds in which the event in question occurs at *some* t, and the chances of detection within a time interval I at each of these worlds are given by the means described above. Although each possible world to which the chances apply is a world in which the event occurs at a definite time, those times are inadmissible before the experiment begins (which is when these probabilities are assigned). The usefulness of 'possible worlds talk' here may suggest that these events could be characterized within an Everettian 'no collapse' interpretation of quantum mechanics, perhaps using the Lewis-friendly possible world semantics of Wilson (2012). This would be consistent with Rovelli's seemingly paradoxical claim that "the statement that a certain specific outcome 'has happened' can be true and not-true at the same time" (2005, p. 115). However, the modern Everettian regards the (world-bound) occurrence of an event as the result of a process of decoherence, no mention of which has been made here. Instead, the occurrence of

these events may be regarded as an indeterministic stochastic process confined to a single world, which is presumably how Haag intends his Principle of Random Realization to be interpreted.

Appendix The proof of Proposition 1 makes use of the following lemma, due to Hegerfeldt (1998).

Lemma 1. *(Hegerfeldt) For any positive operator P, any vector $\psi \in \mathcal{H}$, and any unitary group $U_t = e^{iHt}$ generated by a self-adjoint Hamiltonian H whose spectrum is semi-bounded* either*:*

1. $\langle\psi|U_{-t}PU_t\psi\rangle = 0$ *for all t,* or
2. $\langle\psi|U_{-t}PU_t\psi\rangle \neq 0$ *for (almost) all t.*

Proof. (Of Proposition 1). Let P_c be the projector onto the orthogonal complement of P. At each time $t \in \{t_k\}$ we have $\langle\psi|U_{-t}P_cU_t\psi\rangle = 0$. The premises of Hegerfeldt's Lemma are satisfied by ψ, U_t and P_c. Therefore, $\langle\psi|U_{-t}P_cU_t\psi\rangle = 0$ for *all* t, unless $\{t_k\}$ is a set of zero Lebesgue measure. Assuming that $\{t_k\}$ has non-zero Lebesgue measure, it follows that $\langle\psi|U_{-t}PU_t\psi\rangle = 1$ for all t. Thus $PU_t\psi = U_t\psi$ for all $t \in \mathbb{R}$. Therefore, if ψ is in the range of $P(\{t_k\})$ then ψ is in the range of $P(\mathbb{R})$, i.e. $P(\mathbb{R}) \geq P(\{t_k\})$. But, by definition, if ψ is in the range $P(\mathbb{R})$ then ψ is in the range of $P(\{t_k\})$, i.e. $P(\{t_k\}) \geq P(\mathbb{R})$. Thus $P(\{t_k\}) = P(\mathbb{R})$. □

Inspired by Naimark & Fomin (1957), we define the *extended* Hilbert space $\mathcal{H}_+$ as a continuous direct sum of instantaneous Hilbert spaces $\mathcal{H}_t$, each with inner product

$$\langle\Phi|\Psi\rangle_t = \sum_k \langle\phi(t)|e_k\rangle\langle e_k|\psi(t)\rangle,$$

where $\{e_k\}$ is a fixed orthonormal basis for $\mathcal{H}$. A function $\Psi(t)$ is measurable if $f(t) = \langle\phi|\psi(t)\rangle$ is measurable (with respect to the usual Borel measure on $\mathbb{R}$) for all $\phi \in \mathcal{H}$. If two such functions $\Psi(t), \Phi(t)$ are measurable, then so is the numerical function of t defined by their instantaneous

inner product $F(t) = \langle\Psi|\Phi\rangle_t$. The set of all such measurable functions is a Hilbert space,[9] which corresponds to the continuous direct sum of the spaces $\mathcal{H}_t$, that is, an integral with respect to Lebesgue measure:

$$\mathcal{H}_+ := \int_{\mathbb{R}} \oplus \mathcal{H}_t \, d\sigma(\mathbb{R}).$$

The inner product on $\mathcal{H}_+$ may now be defined as

$$\langle\Phi|\Psi\rangle_+ = \int_{\mathbb{R}} \langle\Phi|\Psi\rangle_t \, d\sigma(\mathbb{R}). \tag{1}$$

It may be verified that $\mathcal{H}_+$ is thus a Hilbert space, and the condition for inclusion of a function Ψ in $\mathcal{H}_+$ is $\langle\Psi|\Psi\rangle_+ < \infty$. This means that $\Psi(t) = U_t\psi$ with $t \in \mathbb{R}$ is not included in this space, but *partial* dynamical evolutions of the system are, i.e. if $\Psi(t) = U_t\psi$ for $t \in [t_1, t_2]$, 0 otherwise, then $\Psi \in \mathcal{H}_+$.

References

Brunetti, R. & Fredenhagen, K. (2002). Time of occurrence observable in quantum mechanics. *Physical Review A*, *66*(4), 044101.

Haag, R. (2013). On the sharpness of localization of individual events in space and time. *Foundations of Physics*, *43*(11), 1295–1313.

Hegerfeldt, G. & Muga, J. (2010). Symmetries and time operators. *Journal of Physics A: Mathematical and Theoretical*, *43*(50), 505303.

Hegerfeldt, G. C. (1998). Causality, particle localization and positivity of the energy. In *Irreversibility and Causality: Semigroups and Rigged Hilbert Spaces* (pp. 238–245). Springer.

[9]Identifying, as usual, functions that differ only on a set of measure zero.

Lewis, D. (1981). A subjectivist's guide to objective chance. *Studies in Inductive Logic and Probability*, *2*, 267–297.

Lewis, D. K. (1986). Events. In *Philosophical Papers II*. Oxford University Press.

Malament, D. B. (1996). In defense of dogma: Why there cannot be a relativistic quantum mechanics of (localizable) particles. In *Perspectives on quantum reality* (pp. 1–10). Springer.

Misra, B. & Sudarshan, E. C. G. (1977). The Zeno's paradox in quantum theory. *Journal of Mathematical Physics*, *18*(4), 756–763.

Naimark, M. & Fomin, S. (1957). Continuous direct sums of Hilbert spaces and some of their applications. *Am. Math. Soc. Translations*, *5*, 35–66.

Rovelli, C. (2005). Relational quantum mechanics. In *Quo Vadis Quantum Mechanics?* (pp. 113–120). Springer.

Wightman, A. S. (1962). On the localizability of quantum mechanical systems. *Reviews of Modern Physics*, *34*, 845–872.

Wilson, A. (2012). Everettian quantum mechanics without branching time. *Synthese*, *188*(1), 67–84.

Is Organismic Fitness at the Basis of Evolutionary Theory?

Charles H. Pence[1] and Grant Ramsey[2]

Fitness is a central theoretical concept in evolutionary theory. Despite its importance, much debate has occurred over how to conceptualize and formalize fitness. One point of debate concerns the roles of organismic and trait fitness. In a recent addition to this debate, Elliott Sober argues that trait fitness is the central fitness concept, and that organismic fitness is of little value. In this paper, by contrast, we argue that it is organismic fitness that lies at the bases of both the conceptual role of fitness, as well as its role as a measure of evolutionary dynamics.

1. Introduction. In a recent paper, Elliott Sober (2013) argues that the fitness of individual organisms in the sense usually described by propensity theorists is urtseless to the actual practice of evolutionary biology. Rather, the crucial sense of fitness for the study of evolution is the fitness of traits, and it is "population-level variation in [trait] fitness" – rather than the absolute value of trait fitness – "that is a causal propensity" (p. 337). Indeed, Sober argues that only for variations in trait fitness can a tenable propensity interpretation be constructed; there exists no consistent propensity account of trait fitnesses themselves.

Sober's argument has much to recommend it. First and foremost, his clarity regarding the distinction between individual fitness[3] and the fitness of traits, as well as the relationship between the

[1] Department of Philosophy and Religious Studies, Louisiana State University

[2] Department of Philosophy, University of Notre Dame

[3] While the debate over biological individuality and the levels of selection is undeniably relevant to work on the concept of fitness (Bouchard and Huneman 2013), the term 'individual' should be

two, has been sadly lacking in recent literature on fitness.[4] But we will argue here that his central thesis – that individual fitness is broadly irrelevant – is mistaken, and that this mistake arises as a result of confusion over the variety of roles that the notion of fitness plays in evolutionary theory. While trait fitness is the salient concept for some of the roles of fitness, for other uses – and uses in which philosophers are particularly interested – it is individual fitness that is the relevant fitness concept. Sober's conclusion is thus too hasty; individual fitness remains vital to the practice of evolutionary biology, and for the interpretation of evolutionary theory.

Many of the most important uses of fitness fall under two categories. First is what we will call a *metrological* role of fitness – that is, fitness's role as a quantitative measure in evolutionary studies. Biologists can measure the realized fitness of organisms by tallying such things as their lifetime reproductive success, and they can measure trait fitness by recording trait changes over time.

Second is what we will call the *conceptual* role of fitness – that is, fitness as an element of the causal or explanatory structure of evolutionary theory. It is this sense of fitness to which Abrams appeals when he says that "the kind of fitness relevant to natural selection is fitness of *types,* that is, properties of organisms, since it is types that are heritable and selected for" (2009, pp. 751-752), and to which Pence and Ramsey appeal when they argue that "organismic fitness plays important roles in

taken to be equivalent to 'organism' in the following.

[4] In his paper, Sober refers to organismic fitness as "token" fitness, while trait fitness is referred to as "type" fitness. We avoid these locutions for several reasons. First, one could construct "type-organism" concepts of fitness. Second, while traits are something like "types" in the sense familiar from metaphysics, they are restricted to particular populations and environments (that is, we are not interested in the fitness of the *type* "organism with brown fur," but in the fitness of the trait "brown fur" within a population, in an environment, at a given time). To avoid these (and other) complications, we will refer exclusively to trait and organismic (or individual) fitness in the following.

parts of ecology and evolutionary biology, and is the concept of fitness underlying the [propensity interpretation of fitness]" (2013, pp. 871-872). Here we are considering a deeper, interpretive question about natural selection – fitness either plays some sort of causal or explanatory role in the theory of evolution by natural selection or it does not – and if it does play a role, then the specifics of that role need to be clarified. It is this role of fitness that we refer to as its conceptual usage.

Keeping this distinction in mind, then, our argument proceeds as follows. In section 2, we argue that there exist three common conceptions of trait fitness – and each of these, in turn, is parasitic on individual fitness, making individual fitness the fundamental notion of fitness in the conceptual role. In section 3, we argue that in the metrological role, the situation is less clear – there are certainly studies in which trait fitness is the more important concept. But it is, we claim, far from true that, as Sober argues, "evolutionary biology has little use for [individual] fitness" (2013, p. 336). In a wide variety of examples, we argue, it is indeed the fitness of individual organisms that biologists look to measure, even when they make inferences about the fitness of traits from those measurements. Individual fitness is therefore *fundamental* in the conceptual role, and *useful* in the metrological role, and should thus, *contra* Sober, by no means be rejected outright.

2. The conceptual role of organismic fitness. In order to understand the conceptual role of organismic fitness in evolutionary theory, we must know what trait fitness is and how it is related to organismic fitness. We will therefore begin by reviewing the uses of trait fitness in its conceptual role in the philosophical literature. We will then show how these concepts are related to one another and to organismic fitness, finally arguing that organismic fitness lies at the conceptual basis of each of the trait fitness concepts, and is therefore at the conceptual basis of the theory of evolution by natural selection.

2.1. Three concepts of trait fitness. We will introduce three definitions intended to capture the core conceptual usage of trait fitness. Nothing in this section should, notably, strike philosophers of

biology as particularly surprising or controversial, since these three definitions of trait fitness appear throughout philosophical work on fitness and natural selection.[5] Further, and importantly, as we will note at the end of this section, these three definitions are often interchanged with one another. Despite the fact that these definitions are often treated as terminological variants, we suggest that they are in fact in profound tension, and their being used interchangeably is deeply problematic.

The first concept of trait fitness holds that the fitness of a trait is the average of the fitness values of the individuals that carry the trait.

> (TF1) The fitness of a trait *t* is equal to the average individual (organismic) fitness values of individuals bearing *t*.

Commitment to (TF1) is widespread and quite explicit. To take one example, Sober notes in his (2001) a tendency for equivocation between individual and trait fitness. He then asserts, however, that the choice of trait or individual fitness is merely semantic, because the two are related by (TF1). That is, "the fitness value of a trait is the average of the fitness values of the individuals that have the trait" (2001, p. 26). Many other authors also explicitly adopt this definition, including Mills and Beatty (1979, p. 276), Walsh, Lewens, and Ariew (2002, p. 462), Abrams (2009, p. 752), and Godfrey-Smith (2009, p. 21).

Second, spurred by the usage of fitness within population genetics, trait fitness is often definitionally linked to trait dynamics:

> (TF2) The fitness value of a trait is a quantity that is, given some model of population dynamics, predictive of the future dynamics of that trait in a population.

[5] Notably, they also appear throughout – and are used on both sides – of the debate between "causal" and "statistical" interpretations of evolutionary theory. We do not intend anything here to privilege or argue for one of these two positions over the other; these definitions could describe either causally potent or causally impotent concepts.

This finally lets us cash out some of the value of trait fitness. We want trait fitness to enable us to predict that, in a given population, the fitter traits will, all other things equal, tend to drive out the less fit.

In biological terms, (TF2) is nebulous, since "future trait dynamics" is a multivalent concept. There are countless models connecting fitness to future outcomes, and there are countless future outcomes we might want to observe, from the simple fraction of a trait in a population to times to extinction or fixation. For our purposes, we intend (TF2) not to pick out any one of these as privileged, but as a highly general definition of trait fitness: whatever we might think that trait fitness is, it must give us some (reasonably accurate) handle on future trait dynamics. Consider, for example, the way in which trait fitness is defined in the population genetics literature. In the simplest models of population genetics – haploid organisms reproducing asexually in discrete time without overlapping generations – the "Darwinian fitness," *w,* may directly provide us with the future proportion at some time *t,* of two competing alleles in a population, p_t / q_t, given their initial proportion (Hartl and Clark 1997, p. 215):

$$p_t / q_t = w^t \cdot p_0 / q_0 \qquad (1)$$

In this and many other models of population genetics, the Darwinian fitness is effectively definitionally connected to the changes in allele frequencies over generational time. (TF2), when used by philosophers, seems to capture their concern for preserving this usage of fitness in population genetics.

If (TF2) does not hold, it is often argued, there is no reason to bother with trait fitness in the first place. A good example here is the work of Ariew and Ernst, who argue that we "employ the concept of fitness when we want to explain why a trait spreads through a population when it does," and that it is a condition of the adequacy of a fitness concept that it "enable us to compare the degree to which natural selection will favor the spread of one trait over another, alternative trait" (2009, p. 290).[6]

The third concept of trait fitness invokes fitness's colloquial usage as a description of the

[6] Explicit mentions of (TF2) also appear in Abrams (2009, p. 752) and Krimbas (2004, p. 188).

"advantage" or "benefit" that an individual organism receives in virtue of possessing a trait:

> (TF3) Trait fitness is the reproductive advantage to the individual conferred by possessing the trait.

This definition echoes the original usage of 'fitness' in evolutionary theory – the fact that organisms bearing some traits are "better fitted" to their environment than those with other traits (Darwin 1859).

2.2. The relationship between (TF1), (TF2), and (TF3). Before we consider the relationship between organismic fitness and (TF1)-(TF3), we will briefly consider the relationship between these trait fitness concepts. These definitions are often conflated in the literature, and our analysis here shows that such conflations are deeply problematic.

Consider the pictures of trait fitness invoked by (TF1) and (TF2). If (TF1) is the operative definition of trait fitness, then trait fitnesses, taken to be averages of individual fitness values, are just one of the causal influences responsible for determining future trait frequencies. But now turn to the case of (TF2). If a model like Equation (1) defines trait fitness, then trait fitness *includes* the effect of (at least) heritability – future trait frequencies are determined *only* by current trait frequencies and current trait fitnesses. Trait fitness in the sense of (TF1) does not include the impact of heritability, but trait fitness as (TF2) does. In many populations, therefore, (TF1) and (TF2) will result in different values for the fitnesses of traits.

The same argument applies to the relationship between (TF2) and (TF3). If a trait has a significant benefit to individual organisms, yet is not (or not efficiently) transmitted from parents to offspring, then the (TF3)-fitness of that trait may be high while its (TF2)-fitness remains low.

Finally, the relationship between (TF1) and (TF3) is similarly complex. Consider a trait which constitutes a fairly minor benefit to organisms, and the (TF3)-fitness of which is hence relatively small. If this trait were to occur only in organisms possessing an otherwise extremely fit genetic background,

then the (TF1)-fitness of the trait might nonetheless be quite high. As another example, a novel trait could be instantiated in a sterile individual. In such a case, this trait would have a (TF2)-fitness of zero, as the only individual organism bearing it will have no offspring whatsoever, and hence has an individual fitness value of zero. And this would be true regardless of the trait's (TF3)-fitness value. The average fitness of the individuals bearing a trait can be large (or small), that is, *without* the effect on individuals being positive and large (or negative and deleterious) in all cases.[7]

It is also noteworthy that the *ranges of possible values* for these different notions of trait fitness differ.[8] Individual fitness values can only be positive numbers (an individual cannot have negative fitness), so the (TF1)-fitness of a trait can only be positive. The (TF2)-fitness or (TF3)-fitness of a trait, on the other hand, can clearly be negative – if a trait is declining in frequency within a population, or if it is deleterious to the individual which holds it, then its (TF2)- or (TF3)-fitness values, respectively, will be less than zero.

2.3. The relationship between trait and individual fitness. It is clear, due both to the extensive use of trait fitness in the literature and the wide variety of ways in which it is defined, that Sober is quite right to argue that trait fitness is an important component of the conceptual foundations of evolutionary theory. But, as noted in the introduction, we take issue with his claim that trait fitness is the *conceptually fundamental* notion of fitness in evolutionary theory. We will now demonstrate that, for each of the three definitions we offered of trait fitness above, *organismic fitness* is the conceptually fundamental concept. While trait fitness concepts are valuable, individual fitness serves as the

[7] Further instances of this sort can be constructed by appealing to the effects of variance on fitness, as described by Gillespie (1974), or by considering cases of pleiotropy – a pleiotropic trait can have only one (TF1)-fitness (the average of its varying effects on organisms with different genetic backgrounds), but its (TF3)-fitness might vary radically across those different organisms.

[8] Normalizing these values could, of course, solve this, but this approach is not taken in the literature.

conceptual foundation for all our uses of fitness in evolutionary theory.

Consider first (TF3). In order to properly apply (TF3) to a particular trait, we need to have a grasp on the appropriate notion of "benefit to the individual."[9] How are we to understand such a concept? As mentioned above, many possible "benefits" can be conceived. They all share one important characteristic in common, however – all will involve references to the fitness of individual organisms. Precisely the challenge of developing a model of individual fitness is to determine the way in which various putatively beneficial influences should be factored into the overall picture offered by fitness. Importantly, though, it is precisely this work that needs to be performed in order to clarify the notion of "benefit to the individual" that is invoked by (TF3). To put the point differently, the work of fully specifying (TF3) to the extent that it can actually be used to describe any particular trait will require the construction of a measure of benefit to the individual. This, in turn, *just is* the construction of a model of individual fitness. However (TF3)'s invocation of benefit might be cashed out, then, it will ultimately depend on some concept of individual fitness.[10]

The conceptual dependence of (TF1) on individual fitness is nearly trivial – if trait fitness simply is the average of individual fitness values, then individual fitness is assuredly the conceptually fundamental notion for (TF1). On (TF1), trait fitnesses can be defined in terms of individual fitnesses, but the converse is impossible. Similarly, information about individual fitness can derive (TF2) values, but (TF2) values cannot derive individual fitness values.

---

[9] One could, conceivably, have a "type"-based notion of (TF3), where the discussion of "benefit" was of "benefit to the type" (thanks to Elliott Sober for pointing out this possibility). It is not clear to us, though, that this would resolve the issues raised here: it does not seem plausible that one could somehow discover what the benefit to a *type* of individual is without clarifying the benefit to *token* individuals.

[10] For a discussion of some of the problems that models of individual fitness need to overcome, see Sober (2001), Abrams (2009), and Pence and Ramsey (2013).

The most difficult case is (TF2). As Sober noted, (TF2)-fitness is in fact a fairly heterogeneous property, including such effects as heritability and individual fitness. The question at hand is whether, like for (TF1) and (TF3), individual fitness also lies at the conceptual basis of (TF2). We contend that this is indeed the case. Our argument for this conclusion is that when (TF2) is analyzed, individual fitness is one of its core components, but not vice versa. To see this, consider that (TF2) is a rate of change in a population. If we ask what underlies this rate of change, the answer will involve several components. If there is immigration, than the immigrants can change trait frequencies. Similarly, emigration can change frequencies, especially if there is a difference in the propensity of different types in the population to emigrate. Mutations and transmission biases, though often small effects, can also change population trait frequencies. All of these factors can change the way in which natural selection operates – but none of them *is* natural selection, and one of the main causes of trait frequency change (or stability) remains the individual fitness values of the organisms in the population. Although there can be (TF2) values in the absence of individual fitness differences, such (TF2) values would not indicate an adaptive response. Instead, they are merely due to migration, mutation, drift, and so forth. It is thus true that when we analyze (TF2), organismic fitness is not just an important factor, but *the* central factor for understanding the adaptive import of (TF2) values.

Now consider individual fitness. Is (TF2) at its basis? The answer is no: Individuals have fitness values that help lead to (TF2) values, but because (TF2) takes into account population factors like drift and migration, and because such factors are extrinsic to the propensities of individuals to survive and reproduce, there is no sense in which (TF2) lies at the conceptual foundation of organismic fitness. While it is true that organisms are built out of traits, and it is these traits that crucially determine organismic fitness values, it is not true that trait *fitness* determines organismic fitness values. (TF2) and organismic fitness clearly bear an asymmetric relation to one another, and it is organismic fitness that is conceptually primary.

We should pause here to deal with one objection. A response to this discussion of (TF2) might

run as follows: Of course (TF2) is not a complete account of the fitness of traits – we need to include explicit accounts of other properties, such as heritability, population/trait dynamics, and so forth. Once enough of these factors have been considered, only then can we say that we've arrived at a true account of trait fitness.[11] Our reply to this objection is that it seems to invoke something like a limiting process, where we begin with the limited information offered to us by (TF2) and add to it until we have arrived at a "complete" picture. But in what would this complete picture consist? It seems, we claim, that some notion like the concept of "benefit to the individual" invoked by (TF3) must be the "target" of the limit, and this would therefore collapse a (TF2)-notion of trait fitness into one based on (TF3). In this case, all the arguments that we deploy with (TF3) would then apply.

It is thus clear that, however we choose to define trait fitness, we are left with a notion of trait fitness that fundamentally depends on the concept of individual fitness. As far as the *conceptual* role of trait fitness is concerned, then, it is the case that individual fitness *always* stands conceptually prior to trait fitness.

Of course, as mentioned above, the conceptual role is not the only one in which trait fitness features. When Sober argues that "*biologists don't bother* with the fitness of Charlie the Tuna, though they may want to discuss the fitness of tuna dorsal fins" (2013, p. 337, emphasis added), he presumably means that individual fitnesses are of little-to-no use in the empirical arena, or for what we called the metrological role of trait fitness. It is to this role that we now turn.

3. The metrological role of organismic fitness. At first blush, it would seem that Sober's argument against the usefulness of organismic fitness rests on entirely plausible premises. The fitness of organisms is typically inaccessible. This is because "organisms taste of life but once" (Sober 2013, 337). Sober's argument seems to say that even though organisms have fitness values, unless the values are zero (through infertility, say), we cannot measure them. We saw in the first section that this

[11] Thanks to Elliott Sober for offering this response.

measurement-focused (*metrological*) role of individual fitness can be distinguished from its conceptual role. Because of this, individual fitness can clearly be the conceptual foundation of evolutionary theory even if it is not readily measurable. In this section, however, we would like to address the metrological question. Is it really true that biologists never care about or measure the fitness of Charlie the Tuna?

One excellent resource for gauging the degree to which individual fitness plays a role in evolutionary studies comes from Endler's (1986) classic monograph on the study of natural selection in the wild. In Chapter 3, Endler identifies ten distinct methods for studying natural selection in the wild. These methods vary from method I, which seeks correlations between environmental factors and traits, to method X, which compares optimization models with actual trait distributions. It is clear that for some of the methods, it is traits that are central, not individual organisms and their fitness values. But for at least some of the methods, the fitness of individual organisms plays a central role. Consider method VII, cohort analysis. In Endler's words, "By gathering detailed data on individuals, data can be obtained on survivorship, fertility, fecundity, mating ability, and so on. Data on parents and offspring can also provide information on genetics (condition *c* for natural selection, inheritance). Data are best gathered from individually marked individuals, though some information can be gained by giving all members of the same cohort the same mark" (1986, 81). This method clearly focuses on individual fitness. But in order for method VII to serve as a counterexample to Sober, we will need some sense of how often this method is used in studies of natural selection in the wild.

Method VII is not one that is easy to perform, especially for some taxa. As Endler notes, it "can be the most laborious method" (81). Does the fact that it is this laborious, however, mean that it is so useless, that, as Sober argues, biologists need not (or cannot) bother with attempting to measure individual fitness values? Fortunately for us, Endler took the trouble to conduct a through survey of studies directly demonstrating natural selection in the wild. His Table 5.1 lists 139 species along with the methods used in the study of each species as well as the publications that have described these studies. If Sober is right that individual fitness is worthless, we should find that few or none of the

studies listed in the table employ method VII. It turns out, however, that a majority of the species (~57%) listed in the table have had natural selection demonstrated in populations via method VII, that is, 79 species mention VII as a method in their studies. Method VIII, which also sometimes focuses on individual fitness (though combined together into "age classes" of individuals), is mentioned for 57 species. If we subtract the 18 species whose study has involved both VII and VIII, we have a total of 118 species that have been subject to methods VII or VIII, 85% of the total. Thus if we assume that Endler's list is representative of the kind of studies conducted today, we can't avoid the conclusion that individual fitness dominates the metrological role of fitness.

On the face of it, then, it seems that biologists wishing to demonstrate natural selection in the wild do care about the fitness of individuals. Charlie the Tuna's fitness is worth measuring, after all. In the previous section, we showed that individual fitness is at the conceptual foundation of evolutionary theory, and in this section we have shown that individual fitness plays important metrological roles in many (or even most) evolutionary studies. The claim that individual fitness is useless, then, is difficult to maintain. Is the case closed on individual fitness being the metrological and conceptual foundation of evolutionary biology? Before we can draw this conclusion, we should consider a case study, focusing on just what sort of role individual fitness actually plays in evolutionary studies of the type that Endler catalogued.

Consider a typical method VII study, that of Booth (1995). Booth tattooed damselfish in a reef ecosystem and then tracked their fates. By following the outcomes of individual life histories, the study was centered on individual fitness. The determination of individual fitness was not, however, the aim of the study. Rather, Booth was trying to determine the impact of grouping behavior on individual fitness. Is it a fitness advantage to be prone to join groups? And are larger or smaller groups the best ones to join? In terms of our (TF1)-(TF3) framework, we can understand the study as proceeding this way: Individuals are identified and their fitness values are recorded along with traits of interest (in this case the characters of the groups they belong to). The data from similar individuals can then be averaged,

resulting in the (TF1)-fitness of the traits measured. This average was then used to parameterize models that offered predictions about future evolutionary dynamics (TF2), and also to estimate the impact that various group sizes have on the individual (TF3).

Thus, just as biologists will be more interested in how dorsal fins affect tuna fitness than the fitness of an individual tuna, they will also be more interested in the fitness effect of particular traits (like tending to join large groups), than in the fitness of particular damselfish. In such cases, individual fitness is frequently used as a means of exploring questions about the evolution of traits. But even if this is true, it still does not mean that individual fitness does not play an important role. In fact, we hope to have shown that the fitness of individuals serves as the basis for the demonstration of natural selection in a large percentage of these kinds of empirical studies. This is perhaps to be expected if, as we argued in section 2, individual fitness lies at the conceptual basis of evolutionary theory.

4. Conclusion. We have argued against Sober's contention that individual fitness is useless to the practice of evolutionary biology. While we agree that trait fitness is sometimes the biologist's sole focus, two facts make Sober's claim incorrect. First, conceptually, each of the three common definitions of trait fitness in fact conceptually relies upon the fitness of individual organisms. Organismic fitness thus lies at the *conceptual* basis of trait fitness. And second, even when biologists are attempting to measure the fitness of traits, they often do so in ways that rely, either tacitly or explicitly, on organismic fitness, making it fundamental as well for the *metrological* role of trait fitness. Organismic fitness, therefore, is crucial to both the theory and practice of evolutionary biology.

References

Abrams, Marshall. 2009. "The Unity of Fitness." *Philosophy of Science* 76 (5): 750–761. doi:10.1086/605788.

Ariew, André, and Zachary Ernst. 2009. "What Fitness Can't Be." *Erkenntnis* 71 (3): 289–301. doi:10.1007/s10670-009-9183-9.

Bouchard, Frédéric, and Philippe Huneman. 2013. *From Groups to Individuals: Evolution and Emerging Individuality.* Cambridge, MA: The MIT Press.

Booth, David J. 1995. "Juvenile Groups in a Coral-Reef Damselfish: Density-Dependent Effects on Individual Fitness and Population Demography." *Ecology* 76 (1): 91–106.

Darwin, Charles. 1859. *On the Origin of Species*. 1st ed. London: John Murray.

Endler, John A. 1986. *Natural Selection in the Wild*. Princeton, NJ: Princeton University Press.

Gillespie, John H. 1974. "Natural Selection for Within-generation Variance in Offspring Number." *Genetics* 76: 601–606.

Godfrey-Smith, Peter. 2009. *Darwinian Populations and Natural Selection*. Oxford: Oxford University Press.

Hartl, Daniel L., and Andrew G. Clark. 1997. *Principles of Population Genetics*. 3rd ed. Sunderland, MA: Sinauer Associates.

Krimbas, Costas B. 2004. "On Fitness." *Biology and Philosophy* 19 (2): 185–203. doi:10.1023/B:BIPH.0000024402.80835.a7.

Mills, Susan K., and John H. Beatty. 1979. "The Propensity Interpretation of Fitness." *Philosophy of Science* 46 (2): 263–286. doi:10.1086/288865.

Pence, Charles H., and Grant Ramsey. 2013. "A New Foundation for the Propensity Interpretation of Fitness." *British Journal for the Philosophy of Science*. doi:10.1093/bjps/axs037.

Scriven, Michael. 1959. "Explanation and Prediction in Evolutionary Theory." *Science* 130 (3374): 477–82.

Sober, Elliott. 2001. "The Two Faces of Fitness." In *Thinking About Evolution: Historical, Philosophical, and Political Perspectives*, edited by Rama S. Singh, 309–321. Cambridge, MA: The MIT Press.

Sober, Elliott. 2013. "Trait Fitness Is Not a Propensity, but Fitness Variation Is." *Studies in History and Philosophy of Biological and Biomedical Sciences*. 44: 336–41. doi:10.1016/j.shpsc.2013.03.002.

Walsh, Denis M., Tim Lewens, and André Ariew. 2002. "The Trials of Life: Natural Selection and Random Drift." *Philosophy of Science* 69 (3): 429–446. doi:10.1086/342454.

Chapter 1

Intensive and Extensive Quantities

(Rough draft as of October 29, 2014. Please do not cite.)

Z. R. Perry

1 Introduction: Physical Quantities

1.1 The Problem of Quantity

Physical quantities—like mass, charge, volume, and length—are commonly represented in science and in everyday practice by mathematical entities, like numbers and vectors. For instance, we use real number and a unit to represent the determinate magnitudes of mass (like $2kg$, $7.5kg$, πg, etc.). These representations are appropriate because they faithfully represent the physical world as being a certain way, as exhibiting certain structural features. Specifically they represent what we might describe as these physical quantities being structured in a certain way.

There has been a long standing problem in explaining exactly what this physical structure consists in. The difficulty lies in giving an account of quantitative structure without either (1) making ineliminable appeal to abstract Platonic mathematical entities themselves (which seem ill suited to explain their own adequacy as representational tools) or (2) positing primitive, irreducible metric structure at the fundamental level (for instance, a distinct and primitive 'n times as long as' relation for every real n).[1] Call this the *problem of quantity*.

[1] I won't motivate these added constraints here. I take it that the motivations for the latter constraint are transparent. An uncountable infinity of distinct primitive posits is the sort of thing that should be avoided wherever possible. Field (1984) makes the best case for the former constraint. I'll just point out that even the most red-blooded Platonist ought to be suspicious of the idea that the numbers 6 and 10 are somehow involved in the ultimate explanation of, e.g., why this $6kg$ ball ricocheted at this particular speed and angle when it collided with this $10kg$ one.

I examine the ways physical quantities constrain the structure of their worldly instances, specifically the mereology of the physical entities which instantiate them. In this paper, I identify a phenomena which I call "proper extensiveness". Of the physical quantities which do put constraints on mereology, including those one might classify as "additive", only a proper sub-class qualify as properly extensive.

In what follows, I will provide motivations for positing such a phenomena, and argue that proper extensiveness cannot be dependent on dynamics. In the second half of the paper, I make the case for taking proper extensiveness to be metaphysically fundamental (at least relative to most of our other physical ontology), by showing that doing so allows us to construct an elegant and attractive solution to the problem of quantity (though only as it applies to quantities which are properly extensive).

Here's the plan for the paper: The rest of this section contains a primer on quantitative structure and establishes some terminology. The argument that we need to posit proper extensiveness is made in sections 2 and 3. Section 2 introduces a puzzle about explaining the reliable success of paradigm physical measurements. The worry is that no explanation that essentially appeals to dynamics can account for the success of synchronic length measurements, like those involving pairs of aligned rods. The best explanation for this success, I argue, requires a pre-dynamical but modally robust connection between quantitative structure and mereology.

Section 3 outlines two candidate connections, one commonly known as "additivity" and the other a previously unrecognized phenomena which I dub "proper extensiveness". I show that only proper extensiveness is sufficient to underwrite the explanation of the length measurement presented in section 2. Also, taking length to be properly extensive better accords with our modal intuitions involving the quantity.

The final section offers a sketch of an application of the distinction to the problem of quantity. I describe a solution to the problem of quantity only available to properly extensive quantities, as I understand them. I discuss the implications such a result would have on our understanding of the nature and significance of proper extensiveness.

1.2 Quantitative Structure

Every physical quantity is associated with a class of determinate *magnitudes* or *values*, each member of which is a (non-quantitative) property or relation itself. So when a particle possesses mass, charge, or length, it always instantiates one particular *magnitude* of that quantity – like $2.5kg$, $7C$ or $2\pi m$.[2] These magnitudes exhibit, or the objects which instantiate them, exhibit "quantitative structure" just in case they are related to one another by certain "structural relations".

We can represent these relations as between magnitudes and between the instances of magnitudes. Some of them are *metrical*—we say "this pumpkin is over 8.7 times as massive as that gourd" when talking about objects and "$1.5m$ is ten times as much as $15cm$" when talking about magnitudes. Other structure is *sub*-metrical. Let me introduce two relations which handily express the sub-metrical structure we intuitively apply to one-dimensional unsigned scalar quantities,[3] i.e. things like mass, length, and volume (and unlike charge, velocity, and spin).

We say "this pumpkin is less massive than that table" and "$22m^3$ is less than $22.1m^3$", when talking about the *ordering* on (in these cases) massive objects and determinate magnitudes of volume, respectively.

Let '$\prec$' denote a two-place relation symbolizing the intuitive "less than" relation over the magnitudes, Q_i, of some quantity, Q. Intuitively $Q_a \prec Q_b$ when Q_a is "lesser than" Q_b. When an object, x, instantiates a mass magnitude that bears $\prec$ to the magnitude instantiated by another object y, we say that x is *less massive than* y.

We say "this stick is as long as that pencil and this highlighter put together" and "$12kg$ is the sum of $7kg$ and $5kg$", when talking about the *summation* or *concatenation* structure on (in these cases) lengthy objects and determinate magnitudes of

[2]It is sometimes said that quantities are determinables and their magnitudes their determinates, but this is not universally accepted. Certainly the magnitudes of mass, say, are all and only the determinates of the determinable property denoted by the predicate 'has mass' or 'has a mass', but it's not obvious that we should identify the *quantity* mass with this determinable property.

[3]By "one-dimensional scalar" quantity, I mean one which is intuitively gradated along only one axis and which don't involve any notion of direction. By an "unsigned" quantity I mean just those which are not most faithfully divided into categories like "positive" and "negative", where two magnitudes might have the same "degree" but differ in "sign". In what follows, I will drop these descriptors, but my focus, for simplicity's sake, will always be on quantities of this type.

mass, respectively.

Let '$\oplus$' denote a three-place relation over the Q_i's which serves to map two magnitudes to a third magnitude which is their "sum". So when $\oplus(Q_a, Q_b, Q_c)$ we say Q_c is the "sum" of Q_a and Q_b, and we write $Q_a \oplus Q_b = Q_c$. When $\oplus$ obtains between three length magnitudes instantiated by objects x, y, and z respectively, we say that z is as long as x and y *taken together*.

I will say a bit more about metrical structure, since it is our target. We'll say the ratio of Q_a to Q_b is intuitively 4.767 to 1 when $4.767 : 1(Q_a, Q_b)$. Since we are construing these only as relations between magnitudes and not between magnitudes and numbers, every distinct ratio must correspond to a distinct 2-place relation.[4]

2 Quantities and the World

The primary way that we gain epistemic access to facts about quantities is by performing measurements. However, measurements are interesting physical processes/procedures[5] in their own right, even putting aside their crucial epistemic role.

For our purposes, a "Q measurement" is a physical procedure performed on certain objects, a and b, (though there needn't be just two) which instantiate magnitudes of a particular quantity, Q. Measurements have a *ready state*, a specification of the state of the measurement apparatus and of a and b relative to that apparatus, as well as a set of possible (mutually incompatible) *outcomes*. Outcomes can include things like different possible positions of a pointer, the relative positions of plates on a balance scale, or a distribution of illuminated pixels on a readout screen.

A measurement's ready state and the different possible outcomes should be distinguishable without appeal to quantitative features of or relations between a and b

[4]This is why doing away with mathematical entities but still positing irreducible metrical structure is an unacceptable solution to the problem of quantity. It requires making an unweildy (indeed, infinite) number of distinct, primitive posits.

[5]I prefer the term 'procedure' to 'process', and will use the former in what follows. This is for two reasons. First, the same procedure can have different outcomes. Second, processes take time, while some measurement procedures are instantaneous (Section 2.2 gives an example). We could think of procedures as event-types which can be tokened in a few importantly different ways, where these differences amount to the different "possible outcomes" discussed below.

(or their respective parts).[6] That is, the ready state of a mass measurement should not include the condition that a be more massive than b, and two possible *distinct* outcomes of a length measurement cannot be distinguished *only* by whether or not a and b bear the "same length as" relation to each other.

Let's call a particular token measurement procedure, performed on a and b, *successful* if the occurrence or non-occurrence of each possible outcome is *reliably correlated* with the obtaining or non-obtaining of different quantitative relations between a and b (or between the magnitudes of Q they instantiate). A successful such measurement procedure produces a counterfactually robust correlation between its outcomes and the quantitative facts—i.e. it renders true conditionals of the form "If a had stood in R_Q to b (at the time of our measurement), then outcome O_i would have occurred".

Such robust correlations, when they obtain, cry out for explanation. A great many such explanations appeal to the role of Q in the dynamics evolving the ready state into one or another possible outcome (I give an example of a mass measurement with such an explanation in Case 1 below). However, certain paradigmatic length measurements do not admit of explanation by such means, yet they still can be robustly successful. Case 2 describes one such successful length measurement, and offers an intuitive, non-dynamical explanation for its success. The rub is, this explanation depends on a substantive connection – which isn't mediated by dynamics! – between length's quantitative structure and the mereology of lengthy physical entities.

2.1 Case 1: Weights on a scale

In the first case, we want to measure the ordering structure (i.e. to determine which, if either, is *more massive* than the other) of a pair of massive objects, a and b.

[6]Indeed, I require that the outcomes of a given measurement procedure must be distinguished wholly non-quantitatively (i.e. *not* by the obtaining or non-obtaining of any quantitative fact, magnitude, or relation). If I was only concerned with the epistemic role of measurement, this last requirement would be needlessly strong. This requirement screens off measurements whose success is really only revelatory of their relationships with and impact on *other quantities*, and not the non-quantitative world directly.

To do this, we set up a balance scale, with two plates suspended from opposite ends of a bar. The bar is balanced at its center on a rigid, vertical stand. The ready state for the scale is with the bar parallel to the ground and weights a and b positioned on opposing plates. We perform the measurement by releasing the plates and waiting a moment or two. The possible outcomes are: a's plate is lower than b's plate, b's plate is lower than a's plate, or the bar is parallel to the ground.[7]

Suppose we run this measurement and get the first outcome—a's plate is lower. Suppose further that a is more massive than b, and that if a had been *less* massive than (just as massive as) b, the second (third) outcome would have obtained. That is, we have performed a *successful* length measurement on a and b. In this particular case, what explains our measurement's success?

Here the explanation should be clear. Mass's quantitative structure plays a certain role in the dynamical laws of motion and gravitation. Specifically, objects which are more massive experience a greater force pulling them towards the earth. After we set the scale up in its ready state,[8] the weights on the scale are impressed by gravitational forces, as dictated by the physical laws. The downward forces on the plates will unbalance a properly calibrated balance scale just in case the objects differ in mass, with the more massive object being pulled more forcefully. Thus the dynamical laws come together with the quantitative facts and the physical makeup of the scale to bring about one of the three outcomes in a way which is reliably correlated with the "less massive than" relation.

Call a measurement procedure of this sort a *dynamical* measurement. Dynamical measurements are successful in virtue of the dynamics governing the evolution from the ready state to the resulting outcome. While there are other ways the dynamics

[7]One might worry that "lower" is a quantitative notion. However, it is not a matter of any quantitative relations between a and b and, in particular, is not a fact about a and b's masses. Even so, this quantitativeness is easy to get rid of, if we complicate our measuring device a bit. Many balance scales have a needle, perpendicular to the horizontal bar, attached at its center. The point of this needle is exactly above the vertical stand when the bar is parallel to the ground, and can either end up still upright or leaning to the left or the right of the vertical stand after.

[8]It turns out that there's some freedom in which ready state you pick. Even if the scale doesn't start with the bar perfectly parallel to the ground, the dynamics on the system will bring it to the right outcome as long as we wait a sufficiently long time.

can be involved in our general measurement *practices*—e.g. in us perceiving which outcome obtains, or in me building a balance scale—a measurement only counts as *dynamical* when the dynamics play an essential role in the measurement's success.[9]

2.2 Case 2: Aligning Rods

We want to measure the ordering structure for a pair of lengthy objects, in this case straight rigid rods. To do this, we arrange the rods so that they are parallel and lay them side-by-side. We then align them at one endpoint—i.e. while keeping them parallel, positioning one endpoint of rod a such that it is immediately adjacent to the endpoint on the same side of rod b. This is the ready state. There are three possible outcomes, as before: rod a extends past rod b, rod b extends past rod a, or neither rod extends past the other (Where "extending past", for these rods, just means one rod having a part which isn't adjacent[10] to any part of the other rod). We observe which of the rods, if either, extends past the other, and conclude that that rod is longer.

Suppose we perform this measurement and get the second outcome—rod b extends past rod a. Let's also suppose that this measurement is successful. I.e. that b is, in fact, longer than a, and that if b hadn't been longer than a, then b would not have extended past a (etc.). What explains the success of *this* measurement procedure?

In this case, we *cannot* appeal to length's role in the dynamics to explain the success of our length measurement. There are, of course, dynamical laws that *involve* spatial quantities like length, but this measurement *has no temporal component*. The procedure's ready state – a and b laid flush against each other and aligned at one endpoint – is *simultaneous* with the procedure's outcome – b's extending past a. Of

[9]Classical mass, it turns out, *only* admits of dynamical measurement. While there are many mass measurement procedures, including various kinds of scales, as well as "collision tests" where massive objects are knocked against each other to see which resists displacement to a greater degree, they all involve an appeal to the dynamics of the measuring system as it evolves from the ready state to one of the resulting outcomes.

[10]We can make the notion of extending past even clearer by doing away with adjacency. For a and b arranged as described in the text, a extends past b just in case there exists a plane orthogonal to a and b which intersects a part of a and no part of b.

course, the dynamics will play a role in our *observing* the outcome after the measurement, and it will play a role in our *positioning* the rods before the measurement, but the dynamics plays no role in evolving the system *from* the ready state *to* the particular outcome. This means that the success of this measurement, and the reliable correlation between its non-quantitative outcome and the quantitative facts, does *not* depend on the dynamics of length or any other quantity.

Indeed, this length measurement could occur and be successful at a world governed by *no* dynamical laws, which exists only for one moment, as long as, at that moment, the rods a and b are situated in the right way.[11]

If not the dynamics, what can explains the success of a length measurement of this sort? This measurement procedure is so transparently legitimate that it's unclear what mechanism *could* be underlying the correlation. The notion of extending beyond is so close to our conception of being longer (or instantiating a length magnitude bearing $\prec$ to the other) that it's hard to see the gap at all, let alone identify what's bridging it. It's not especially difficult to come up with an intuitively satisfying explanation of this case. The trouble is giving an account of what length must be like such that this explanation applies.

[11]There's a bit of nuance here that we should address. The issue isn't *merely* that the ready state and result state are simultaneous, though this is important. The issue is that the connection between them isn't dynamical. For instance, we could construct a mass measurement which could be performed instantaneously, but it would still depend on the dynamics. In 2.1, I pointed out that, in the case of a balance scale, we have some freedom in where we set the angle of the balance bar suspending the two plates, as long as we wait long enough for the system to enter equilibrium.

What makes the outcome of such a measurement important is that it represents an equilibrium state of the system. The state evolves to equilibrium and then stays in that state. Since there's some freedom as to the angle of the bar, we could start with our bar in *exactly the right position* such that the system is *already in an equilibrium state* when we let it go! In this case, there is a certain sense in which the ready state and the outcome are *simultaneous*.

However, the fact that the two states are simultaneous doesn't mean the success of the measurement isn't dependent on mass's role in the dynamics. What gave that outcome its status was that it's an equilibrium state, but being at equilibrium is a dynamical feature. It's a matter of what the dynamics governing that system *would* do to such a system *if* it were left alone and given a chance to evolve. So even if there *were* a world that existed only for an instant and contained a balance scale in exactly the right position, it would only count as a successful mass measurement if there were dynamics "governing" (or that would govern) that short-lived system, and the system was in equilibrium according to those dynamics. A short-lived world *without* any dynamical laws at all could not support such a measurement.

Here's what I think is going on in this case:
b extends past a. So while there's a part of b that is perfectly aligned with a, but there's also a *remainder*—i.e. another part of b that has no part that's adjacent to any part of a. Call the first part x and the second part, the remainder, y. The existence of such parts doesn't yet establish that b is longer than a. For that we need two bridge principles connecting the mereology and the quantitative facts.

(1) If two rods are laid side by side such that neither extends past either endpoint of the other, then they are as long as each other.

(2) A rod must be longer than any of its proper "rod segments".[12]

(1) establishes that a is as long as x. (2) establishes that b is longer than x. Together they establish that, in situations like our length measurement, above, b is longer than a. While premise (1) is of central importance to the practice of measuring length by laying rods side-by-side, I will not be discussing it much here.[13]

[12]Premise (2) is expressed in terms of rules of thumb for measuring rigid rods, and makes use of the notion of a "rod segment". This is not ideal, but it's important to recognize that the more natural sounding principle: "a rod must be longer than any of its proper parts" has some unfortunate exceptions. In particular, a three meter rod could be cut "lengthwise", so to speak, and thus divide into two three meter parts. Alternatively, it also has proper parts that, intuitively, have no length at all, but are just a few spatially disconnected pieces of rod. The notion of "rod segment" is meant to rule out cases like these.

If the reader is still worried that a rod could be as long as one of its "rod segments", perhaps because the rod segment is just the segment of the rod *minus* some lengthless slice at one endpoint, we can add premise (3):

(3) If a rod can be partitioned into two "rod segments", it is longer than each of them.

What premise (3) relies on is the idea that an infinitely thin slice off the endpoint of a rod is not a "rod segment" (even if its complement is). Once we move beyond this example and do away with talk of rods in favor of talk of spatiotemporal *paths*, we can avoid all the ambiguity involved in the notion of a rod segment.

[13]Premise (1) is likely an approximation of a principle that has its source in Euclid, with his Common Notion 4: "Things which coincide with one another are equal to one another."[Euclid (trans. Heath, 1908)]. Since material bodies can't interpenetrate, the closest to coinciding we can come, practically, is alignment without remainder, i.e. being laid side by side with neither extending beyond the other. There's much more to be said about the spatial structure of the world such that this approximation works, to the extent it does, but that's outside the scope of this paper.

2.3 The puzzle of non-dynamical measurement

This is a patently non-dynamical explanation. The outcome (b extending past a) and the quantitative facts (b being longer than a) are correlated, according to this explanation, not because of length's role in the dynamics but because of certain constraints on the mereology of lengthy objects (i.e. on the possible lengths of objects given their mereological structure and relations, and the possible mereological structure of objects given their lengths and length relations.). This connection between quantitative structure and mereology shows up at two points in the explanation:

The first is obvious. Premise (2) establishes that a rod bears a certain quantitative relation (longer than) to every member of a certain special sub-class of its parts.

The second involves premise (1), though in a more nuanced way: The explanation of the success of a length measurement of a given pair of rods, a and b, such that b extends past a, was presented as *fully general*. That is, for *any* rod shorter than b, which is measured against it in this way, b must have a proper part that's perfectly aligned with that rod. By (1) this implies that b has a proper part that's as long as that rod, for *any* such rod shorter than b! Here this explanation (& specifically its generality) puts substantial constraints on the parts of b and the lengths those parts can have.

Before we go any further, we will have to replace this talk of rods with something more rigorous. (1) and (2) are *approximately* true, as is this assumption about the generality of the explanation. However, though the success of the measurement of the rods a and b *can* be *roughly* explained by appeal to these principles, we don't *need* to tether our explanation to the nature of something as derivative and clunky as the notion of a concrete, straight, macroscopic material rod, and the "rod segments" which make it up. Indeed, if we want to give a truly rigorous and completely general explanation, we will need to give it in terms of the fundamental entities and properties in the vicinity.

Let's say that length is, fundamentally, a property of one-dimensional, open (i.e. non-looped) paths through spacetime. To the extent that a concrete material rod can be said to have length, it has its length derivatively, in virtue of occupying

a region containing certain (properly oriented) spatiotemporal paths of such-and-such a length. We should be able to recapture an explanation in terms of rods by appealing to the properties of the regions they occupy. For the remainder of this paper, however, I'll be concerned only with the more general and rigorous principles concerning spatiotemporal paths.

We can capture the significance of premise (2) and of the generality assumption in one principle (By "object of length L_n" I'm only referring to things like substantival paths, and not to anything which has its length derivatively):

(2′) For all objects x of length L_n, and for all lengths $L_m \neq L_n$, x has a proper part of length L_m iff $L_m \prec L_n$.

(2′) puts very strong constraints on the sorts of parts lengthy objects can have, and on the possible lengths those parts can have. Analogously to (2), (2′) implies that a given path is as long or longer than *all* of its lengthy parts. Analogously to the assumption about generality, (2′) implies that a given path of length L_n must have a lengthy proper part corresponding to *every* length property bearing $\prec$ to L_n.

The only explanation for the reliable success of synchronic length measurement on offer requires a principle like (2′). But neither the physical details of the measurement procedure, nor the dynamical laws governing the system, are responsible for conditions like (2′). If this explanation is a good one, then, our theory of quantities like length should be able to account for the truth of (2′) in the relevant situations. To do this, we will have to consider how certain quantities constrain the mereology of their instances. In the next section, I argue that the way quantities are standardly assumed to put constraints on that mereology is insufficient to underwrite this explanation, and I propose an alternative.

3 Constraining the World

In this section, I consider two ways a quantity might put constraints on the mereological structure of its instances. The first is commonly called "additivity", while the second is a hitherto undiscussed phenomena, which I have dubbed "*proper*

extensiveness" (though I will argue that it better captures some of our modal intuitions concerning certain physical quantities). I will show that additivity, properly understood, cannot explain the success of instantaneous length measurements, while proper extensiveness can.

3.1 Additivity

An additive quantity, roughly, is one for which composite objects "inherit" their Q-value (what magnitude of Q they instantiate) from the Q-values of their parts (if they have any). For instance, mass and length are both additive quantities. $2kg$ and $3kg$ stand in '$\oplus$' relation to $5kg$ ($2kg \oplus 3kg = 5kg$). Since mass is additive, composites of massive objects "inherit" their masses from their parts; so the mereological sum of a non-overlapping[14] pair of objects weighing $2kg$ and $3kg$ must weigh $5kg$.[15] The inheritance analogy is a powerful one, as it indicates both the strength and – we shall see – the limitations of this connection.

Put more formally, an additive quantity necessarily satisfies the following conditionals. They hold for any magnitudes, Q_i (of the same additive quantity), that satisfy the antecedent. The mereological relations used are these: '$O(x, y)$' for overlap, '$(x, y)C(z)$' for a three-place composition relation, with the third relatum being the fusion of the first two, and '$P(x, y)$' for parthood.

Additive $\prec$: $(Q_m \prec Q_n) \rightarrow \forall x \forall y((Q_n(x) \wedge Q_m(y)) \rightarrow \neg P(x, y))$

Additive $\oplus$: $(Q_m \oplus Q_n = Q_r) \rightarrow \forall x \forall y \forall z((Q_m(x) \wedge \neg O(x, y) \wedge (x, y)C(z)) \rightarrow (Q_r(z) \leftrightarrow Q_n(y)))$

[14] If we're being really strict about it, the parts may either have no overlap or have only "negligible overlap". What counts as negligible overlap depends on the structure and mereology of the quantity in question. For instance, often negligible overlap might just be overlap which instantiates the "zero magnitude", like $0m$ or $0kg$ or $0cm^3$. However, if one's metaphysics of the relevant quantity does not include a zero magnitude ([Balashov, 1999] takes issue with the very idea of a zero magnitude, albeit for reasons I'm not especially sympathetic to) the notion of negligible overlap must be got at in a different way.

[15] For ease of presentation, I assume mereological universalism. Certain complications would arise if we were to drop this assumption. However, none of the substantive points of the paper depends on it. I also assume that none of the massive objects discussed are spinning.

In the case of mass, **Additive** $\prec$ says that no massive object can have a part which is more massive than it. **Additive** $\oplus$ says that the fusion of any two non-overlapping objects has the "sum" of their respective mass magnitudes as its mass, providing they instantiate mass magnitudes at all. These conditionals (on the assumption that $\oplus$ is commutative) fully specify the mereological significance of additivity. These conditionals are modally[16] robust: Suppose *pumpkin* is a $5kg$ object composed out of non-overlapping parts *body* and *stem*. If we consider a counterfactual scenario in which the only difference is that *stem* is $2kg$s heavier (than it actually is), we readily (often automatically) infer that at this world *pumpkin* is $2kg$s heavier as well. Indeed, its difficult to conceive of a world where only *stem*, but neither *body* nor *pumpkin*, changes its mass.

3.2 Additivity and Measurement

The reason additivity cannot explain the success of synchronic length measurement is well illustrated by the "inheritance" analogy. Additivity says that an object's mass is determined by the masses of its parts. However, **Additive** $\oplus$ and **Additive** $\prec$ are entirely silent on *whether* a given massive object has parts (massive or otherwise). This means that length's additivity cannot itself account for the truth of $(2')$.

Since **Additive** $\prec$ and **Additive** $\oplus$ never imply that a given object *must have* parts of some kind, they're consistent with a pair of objects, a and b, instantiating magnitudes, Q_a and Q_b, (of some additive quantity) where $Q_a \prec Q_b$ yet both a and b are *mereological simples*. There's nothing obviously wrong with admitting of such a possibility for *mass*. On the ordinary understanding of most particle theories, elementary particles are assumed to be mereologically simple, and there is no prohibition on different elementary particles ever possessing different masses! How-

[16]The nature of this modal robustness, i.e. the degree of necessity possessed by the conditionals **Additive** $\oplus$ and **Additive** $\prec$, is not entirely clear, and may differ from quantity to quantity. For instance, on some understandings of classical mass, on which it is identical to inertia, the truth of **Additive** $\oplus$ and **Additive** $\prec$ for mass might be grounded in the nature of mass's dynamical role. If so, then these conditionals may well be merely nomologically necessary, when it comes to mass, rather than metaphysically necessary.

ever, the analogous possibility for *lengthy* entities is flatly inconsistent with (2′).[17] Moreover, such a possibility also seems to get the modality of fundamentally lengthy entities (like spacetime paths or trajectories) intuitively wrong (I go more in depth into this issue in particular in the next section).

Mere additivity cannot explain the reliable and general success of synchronic length measurement.

3.3 Proper Extensiveness

I'm going to introduce a phenomena called "proper extensiveness". My contention is that certain physical quantities are properly extensive—length, volume, and temporal duration among them—and that properly extensive quantities, by their nature, put stronger constraints on the mereological structure of the world than merely additive quantities do. Specifically, these constraints are sufficient to entail (2′) for length and thereby support the intuitive explanation offered in the previous section for the success of synchronic length measurement.

Physical quantities can be grouped into the additive and the non-additive (sometimes called "intensive") quantities, and the class of additive quantities can be further divided into the *merely* additive quantities and the *properly* extensive quantities. As such, properly extensive quantities satisfy **Additive** $\prec$ and **Additive** $\oplus$:

Additive $\prec$**:** $(Q_m \prec Q_n) \rightarrow \forall x \forall y((Q_n(x) \wedge Q_m(y)) \rightarrow \neg P(x,y))$

Additive $\oplus$**:** $(Q_m \oplus Q_n = Q_r) \rightarrow \forall x \forall y \forall z((Q_m(x) \wedge \neg O(x,y) \wedge (x,y)C(z)) \rightarrow (Q_r(z) \leftrightarrow Q_n(y)))$

$2m$ and $3m$ stand in $\oplus$ to $5m$ (i.e. $2m \oplus 3m = 5m$). Length is additive, so the fusion of two non-overlapping objects of length $2m$ and $3m$ laid end-to-end (in the right way) will be $5m$ long. If length were *merely* additive, that would be the end of the story. Because length (we are supposing) is also properly extensive, we can say

[17] To see this, realize that it's also consistent with the dictates of additivity that there be two lengthy objects, a and b, of lengths, $2m$ and $5m$ respectively, where b has no proper part as long as a (that is, $2m$ long) because b is a mereological simple.

more—e.g., since $2m \oplus 3m = 5m$, *any* $5m$ path *must* admit of a partition into a $2m$ part and a $3m$ part. We'll understand a partition of o as a class of non-overlapping objects whose fusion is o. That is, properly extensive quantities also necessarily satisfy:[18]

Extensive $\prec$: $(Q_m \prec Q_n) \Rightarrow \forall x(Q_n(x) \rightarrow \exists y(y \neq x \wedge Q_m(y) \wedge P(y, x)))$

Extensive $\oplus$: $(Q_m \oplus Q_n = Q_r) \Rightarrow \forall x(Q_r(x) \leftrightarrow \exists y \exists z(Q_m(y) \wedge Q_n(z) \wedge \neg O(y, z) \wedge (y, z)C(x)))$

In the case of length,[19] what **Extensive** $\prec$ says is that every spatial path of a given length L_n, such that $L_m \prec L_n$, has an interval (which is to say, a part which is itself a path) of length L_m. **Extensive** $\oplus$ says a path can instantiate a length magnitude L_a such that $L_b \oplus L_c = L_a$, if and *only if* it has two non-overlapping parts which respectively instantiate those magnitudes. This is a very powerful condition, because it says that, given the quantitative facts, just instantiating a given length magnitude, L_a, necessarily requires you to have parts with certian length properties standing in certain mereological relations to one another.

Recall that, in order for our explanation of synchronic length measurement in terms of the existence of a remainder to apply, our theory of length must entail that the quantity satisfies:

[18] Additivity and proper extensiveness both involve principles which concern the quantitative features of objects "put together in the right way". For most quantities, like mass or volume, the formula '$\neg O(x, y) \wedge (x, y)C(z)$' will accurately describe this condition. However, since only certain kinds of objects can have length (namely, unbroken non-looping paths), the conditions for putting two paths together "in the right way" are more stringent. It isn't enough for path a and path b to not overlap and to together compose object c. If a is the spatial path from my nose to my upper lip, and b is the shortest path from the surface of the earth to the moon, then their fusion, c, isn't an unbroken path, and so doesn't have length! The conditions for length would be something like this: a and b are both intervals of path c, which is their mereological fusion, and a and b either don't overlap or have a lengthless overlap (either one with $0m$ length or without length, depending on what we want to say about the lengths of unextended points). Since I am more concerned here with the relationship between the second-order $\prec$ and parthood, I will set this issue aside

[19] Technically these conditionals, as stated, only directly apply to properly extensive quantities like volume or surface area. They would need slight tweaking to accurately characterize a quantity like length. How we sort out this wrinkle won't, however, make a difference for our argument concerning measurement.

(2′) For all objects x of length L_n, and for all lengths $L_m \neq L_n$, x has a proper part of length L_m iff $L_m \prec L_n$.

An account of length on which length is properly extensive does entail (2′). By **Extensive** $\prec$, we get that if $L_m \prec L_n$ then x has a part of length L_m, and by **Additive** $\prec$, we get that if x has a proper part of length L_m, then L_m must be either $= L_n$ or $\prec L_n$ (which, given the assumption that $L_m \neq L_n$, implies that $L_m \prec L_n$.

3.4 The significance of Proper Extensiveness

The fact that proper extensiveness is necessary to explain the reliable success of a paradigm measurement procedure is important because it indicates that (1) there is good reason to take length to be properly extensive and (2) that the necessary conditionals characterizing proper extensiveness must be independent of the operation of the dynamical laws. Solving the puzzle of synchronic length measurement is less an end in itself and more a means to introduce and motivate proper extensiveness. In this section I will further examine this phenomena, and in the next outline a very significant application.

Some of our central intuitions regarding physical quantities like length, volume, and temporal duration—specifically those concerning how the mereological structure of the world reflects the quantitative structure of the properties instantiated at it—already suggest a tacit commitment to something like proper extensiveness for these quantities.

One striking consequence of taking length to be properly extensive illustrates this quite well. Suppose we discover a path through space that had a non-zero length, L_u, but no proper sub-paths (i.e. no proper parts which are paths). According to **Extensive** $\prec$, this implies that there are *no* length magnitudes $\prec L_u$ (except the zero-magnitude, $0m$, if there is such a thing)—meaning that the quantity, length, is *discrete* (best represented by the natural numbers plus zero) and that L_u is its *unit length*.

This result very closely accords with our intuitive expectations about what the

physical world can tell us about quantities like length. We do not hear metaphysicians raise concerns when physicists run together the possibility that there is a smallest non-zero length (alternatively, that the quantity *length* is discrete) with the possibility that there are shortest possible *paths* (alternatively, that *space* is discrete). Indeed, many discussions of length readily use "shorter than" and "as long as a proper sub-interval of" interchangeably. Similar points can be made for area, volume, and temporal duration. The pervasiveness of this line of thought disguises how significant of a metaphysical commitment it amounts to, once we take it seriously. The notion of proper extensiveness is how we should characterize this commitment.

It is important to stress again how these commitments simply do not hold sway for merely additive quantities. Though mass's status as merely additive is not entirely uncontroversial, treating it that way is in accordance with an extremely common understanding of the quantity.[20] On this understanding, there could very well be two simples (objects without proper parts) with differing, non-zero, masses. When entertaining the epistemic possibility that, e.g., the electron is a point-particle (without spatial extension and, it is presumed, mereologically simple), we don't at all expect every *other* elementary particle to therefore be exactly as massive as the electron! However, that is precisely the sort of conclusion we *should* reach in the analogous scenario for quantities like length and volume!

I've suggested that there exists a distinction in our intuitions about the modal mereology of additive physical quantities. If this is right, it stands as strong evidence in favor of a distinction between the additive quantities into the merely additive and the properly extensive, as I draw it. The lack of acknowledgment or discussion of this phenomena in the philosophical and physical literature means that (as of yet)

[20]The fact that mass is closely associated with a certain dynamical role is good evidence that it's not properly extensive, since we standardly think that the same dynamical role in gravitation or inertia could be played equally well by a mereological complex or a simple. However, for all we know it may turn out that mass more closely aligns with earlier notions of physical mass as the "measure of matter". If that is right, to say that a is less massive than b is to say that a has less matter making it up than b. One way to draw out this understanding would be to treat mass as properly extensive, and to expect its instances to obey the associated mereological constraints (i.e. if b has more matter making it up than a does, then b should have a part which has exactly as much matter making it up as a does).

there are no suggestions on the table as to *why* some quantities are extensive, or *how* this constraining of the mereology is supposed to work. For our purposes, it suffices to say *that* some quantities are extensive and *that* they constrain mereology in a modally robust way that is independent of the dynamical laws.

4 Conclusion: Applying Proper Extensiveness to the Problem of Quantity

In the previous two sections I have argued in favor of positing a distinction amongst the additive quantities into the merely additive and the properly extensive. I have argued that this distintion better captures and explains the data, specifically regarding simultaneous length measurement as well as our modal mereological intuitions about various physical quantities. I'd like to close by gesturing in the direction of a significant potential application of this distinction. Specifically, I will give a few reasons to believe that an elegant and principled solution to the problem of quantity, as it applies to properly extensive quantities, is available if we take proper extensiveness as fundamental

Many metaphysicians of quantity appeal to *measurement theory* in their answer to the problem of quantity. Specifically, they attempt to reduce facts about metric structure to facts about the world satisfying the right measurement-theoretic axioms.[21] Measurement theory is a formal discipline which involves rationalizations, formalizations and defenses of empirical measurement practices. The game of measurement theory is to take a domain of material objects, which instantiate different magnitudes of some quantity, Q, posit some axioms that these objects obey, and then prove theorems which imply that Q can be faithfully represented, up to a point, with a certain mathematical structure, e.g. the real numbers.[22]

Some of the axioms required to prove these theorems impose certain requirements on the *size* and *structure* of the domain itself. They say that domains are well

[21] Field (1980) is the most famous account along these lines.
[22] Cf. [Krantz, et. al. 1971]

populated (existence axiom), and that there's ample variation in which magnitudes of Q are instantiated therein (richness axiom). The satisfaction of such axioms is a contingent matter. If there aren't enough objects, or if they don't instantiate enough different magnitudes, these axioms fail to be satisfied.

But our account of the ground of metric structure ought not to be contingent on the world being well-populated! This contingency problem has been acknowledged in the literature, and various theorists have proposed ad hoc solutions to eliminate this contingency. [Mundy, 1987] gives up on the domain of massive *objects* and instead attempts to apply measurement theory to the domain of mass *magnitudes*, while Arntzenius and Dorr in their (2013) avoid the contingency problem by positing well-populated substantival physical spaces, and identifying the geometry of *this* space with the relevant quantitative structure.

The unique advantage of properly extensive quantities is that any world where such a quantity is instantiated must, by the conditions it places on the mereology of its instances, be well populated and variegated, to a certain degree. Suppose that L_x is a length magnitude, instantiated by a path, p. **Extensive** $\prec$ implies that p will have *at least* as many proper parts as there are length magnitudes which bear $\prec$ to L_x. Similarly, **Extensive** $\oplus$ implies that p will admit of a partition into parts of length L_y and L_z, for every such pair of length magnitudes such that $L^y \oplus L_z = L_x$.

This suggests that a domain where a properly extensive quantity is instantiated, and in which its instances satisfy the necessary constraints its proper extensiveness puts on their mereology, may be of the right form to satisfy the relevant existence and richness axioms. I think this can be shown, but there's no room to do so here. However, if it were true, it would allow for a uniquely elegant and principled solution to the problem of quantity, as it applies to properly extensive quantities.

A result of this kind, if it can be done (and I think it can), is not just important because it moves us closer to a satisfactory solution to the problem of quantity in full generality. It also speaks to the metaphysical depth of the distinction between properly extensive quantities and all other physical quantities, one which manifests not just in the way these quantities relate to mereology, but also in the nature and ground of their metric structure.

References

[Mundy, 1987] Mundy, Brent (1987). "The Metaphysics of Quantity". Philosophical Studies 51 (1):29 - 54.

[Eddon, 2013] Eddon, Maya (2013). "Quantitative Properties". Philosophy Compass 8 (7):633-645.

[Eddon, 2007] Eddon, Maya (2007). "Armstrong on Quantities and Resemblance". Philosophical Studies 136 (3):385 - 404.

[Balashov, 1999] Balashov, Yuri (1999). "Zero-value physical quantities." Synthese 119 (3):253-286.

[Bigelow and Pargetter, 1988] John Bigelow , Robert Pargetter & D. M. Armstrong (1988). "Quantities". Philosophical Studies 54 (3):287 - 304.

[Field, 1984] Field, Hartry (1984). "Can We Dispense with Space-Time?" PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1984:33 - 90.

[Field, 1980] Field, Hartry (1980). "Science Without Numbers". Princeton University Press.

[Busse, 2009] Busse, Ralf (2009). "Humean Supervenience, Vectorial Fields, and the Spinning Sphere". Dialectica 63 (4):449-489.

[Krantz, et. al. 1971] David Krantz , Duncan Luce , Patrick Suppes & Amos Tversky (eds.) (1971). *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. New York Academic Press.

[Lewis, 1986] Lewis, David K. (1986). *On the Plurality of Worlds.* Blackwell Publishers.

[Arntzenius and Dorr, 2012] Frank Arntzenius & Cian Dorr (2012). Calculus as Geometry. In Frank Arntzenius (ed.), Space, Time and Stuff. Oxford University Press.

[Euclid (trans. Heath, 1908)] Euclid. *Volume 1 of The Thirteen Books of Euclid's Elements*. trans. Sir Thomas Little Heath, ed. Johan Ludvig Heiberg. The University Press

Aspects of theory-ladenness in data-intensive science[1]

Wolfgang Pietsch[2]
Munich Center for Technology in Society,
TU München, Munich, Germany

Abstract

Recent claims, mainly from computer scientists, concerning a largely automated and model-free data-intensive science have been countered by critical reactions from a number of philosophers of science. The debate suffers from a lack of detail in two respects, regarding (i) the actual methods used in data-intensive science and (ii) the specific ways in which these methods presuppose theoretical assumptions. I examine two widely-used algorithms, classificatory trees and non-parametric regression, and argue that these are theory-laden in an external sense, regarding the framing of research questions, but not in an internal sense concerning the causal structure of the examined phenomenon. With respect to the novelty of data-intensive science, I draw an analogy to exploratory as opposed to theory-directed experimentation.

1. Introduction

Over the past decade, computer scientists have claimed that a new scientific methodology has become possible through advances in information technology (e.g. Gray 2007). This approach is supposed to be data-driven, strongly inductive, and relatively theory-independent. The epistemology of such data-intensive science has recently emerged as a novel topic in philosophy of science. Generally, the reactions have been rather critical, often referring to the more or less trivial observation that some kind of theory-ladenness always occurs in scientific research. But, as I will argue, this means throwing the baby out with the bathwater, since interesting shifts in the role of theory can indeed be observed when examining specific methods employed in data-intensive science.

In Section 2, I will suggest a definition for data-intensive science reflecting those features that are interesting from an epistemological perspective. I will then, in Section 3, briefly introduce the debate on theory-ladenness in data-intensive science. To assess the various arguments, I will discuss two algorithms that are widely used, namely classificatory trees (Section 4) and non-parametric regression (Section 5). For both of these methods, I will point out the specific ways in which theory has to be presupposed to identify causal connections and thus yield reliable predictions. I will conclude in Section 6 that these algorithms require an *external* theory-ladenness concerning the framing of research questions, but little *internal* theory-ladenness concerning the causal structure of the examined phenomena. I will also point out remarkable analogies to the analysis of theory-ladenness in exploratory experimentation.

[1] accepted for the proceedings volume of PSA2014, 24th Biennal Meeting of the Philosophy of Science Association
[2] pietsch@cvl-a.tum.de

2. Defining data-intensive science

The problems usually addressed in data-intensive science bear close resemblance to standard problems in statistics. They concern classification or regression of an output variable y with respect to a large number of input parameters x, also called predictor variables or covariates, on the basis of large training sets. The main differences compared with conventional problems in statistics consist in the high-dimensionality of the input variable and the amount of data available about various configurations or states of the system. For example, an internet store wants to know how likely someone buys a certain product depending on surf history, various cookies and a user profile as well as based on data of other users who have either bought or failed to buy the product. A medical researcher examines which combinations of genetic and environmental factors are responsible for a certain disease. A political adviser is interested how likely a specific individual is going to vote for a certain candidate based on a profile combining for example voting history, political opinions, general demographics, or consumer data.

In a *classification* problem, the output variable has a finite number of possible values. In a *regression* problem, the output variable is continuous. In order to establish an adequate and reliable model, extensive training and test data is needed. Each instance in the training and test sets gives a value for the output variable dependent on at least some of the input parameters. The training data is used to build the model, e.g. determine relevant parameters, the test data to validate and verify the model. Using part of the data to determine the accuracy of a model is commonly referred to as cross-validation.[3]

In this essay, we cannot delve into all the technical details of the various algorithms employed in data-intensive science, such as support vector machines, forests or neural networks. Instead we will look at two simple but widely-used algorithms, namely classificatory trees and non-parametric regression, to examine how much and what kind of theory must be presupposed in order for these algorithms to yield meaningful results.

The term *data-intensive science* is notoriously blurry, as has been emphasized for example by Sabina Leonelli: ‘a general characterisation of data-driven methods is hard to achieve, given the wide range of activities and epistemic goals currently subsumed under this heading.’ (2012, 1) However, in order to say something substantial about the role of theory, we have to be more specific about the kinds of practices we want to include as data-intensive science even if an exact definition does not fully correspond to common usage of the term.

In the computer science literature, various definitions have been proposed for the closely related concepts of a *data deluge* or of *big data*. Most of these refer to the pure amount of information or to the technical challenges that such ‘big data’ poses in terms of the so-called ‘three Vs’—volume, velocity and variety of data (Laney 2001). However, from a philosophy of science perspective, these definitions do not provide much insight. After all, larger amounts of data do not automatically imply interesting methodological developments.

[3] An excellent introductory textbook is Russell & Norvig (2009).

Leonelli, partly following Gray (2007, xix), identifies two characteristic features for data-intensive methodology: 'one is the intuition that induction from existing data is being vindicated as a crucial form of scientific inference, which can guide and inform experimental research; and the other is the central role of machines, and thus of automated reasoning, in extracting meaningful patterns from data.' (2012, 1) She adds that these features are themselves quite controversial and criticizes that they are difficult to apply in research contexts.

In defining data-intensive science, I largely follow Leonelli, while attempting to be more precise about the type of induction. I will argue that eliminative induction in the tradition of Mill's methods[4] plays the crucial role. The first part of my definition thus focusses on the premises that are necessary to carry out eliminative induction: data-intensive science requires (I) *data representing all relevant configurations of the examined phenomenon with respect to a specific research question*. For complex phenomena, this implies high-dimensional data, i.e. data sets involving many parameters, as well as a large number of observations or instances covering a wide range of combinations of these parameters. We will see later that this premise underwrites the characteristic data-driven and inductive nature of data-intensive science.

(II) The second feature concerns the *automation of the entire scientific process*, from data capture to processing to modeling (cp. Gray 2007, xix). This allows sidestepping some of the limitations of the human cognitive apparatus but also leads to a loss in human understanding regarding the results of data-intensive science. Again, being more precise about the type of induction allows to determine under which circumstances automation is really possible.

3. Theory-free science?

Proponents of data-intensive science claim that important changes are happening with respect to the role of theory. An extreme, but highly influential version of such a statement is by the former editor-in-chief of *Wired* Chris Anderson, who notoriously proclaimed 'the end of theory' altogether (2008). More nuanced positions can be found for example in the writings of Google research director Peter Norvig (2009): 'Having more data, and more ways to process it, means that we can develop different kinds of theories and models.' Simpler models with a lot of data supposedly trump more elaborate models with less data (Halevy et al. 2009, 9).

A number of philosophers have objected to claims of a theory-free science—generally by pointing out various kinds of theory-ladenness. For example, Werner Callebaut writes: 'We know from Kuhn, Feyerabend, and […] Popper that observations (facts, data) are theory-laden. Popper […] rejected the "bucket theory of knowledge" in favor of the "searchlight theory," according to which observation "is a process in which we play an intensely active part." Our perceptions are always preceded by interests, questions, or expectations—in short, by something "speculative".' (Callebaut 2012, 74) Leonelli concurs in her work on big data biology: 'Using data for the purposes of discovery can happen in a variety of ways, and involves a complex ensemble of skills and methodological components. Inferential reasoning

[4] not to be confused with a looser use of the same term in the sense of eliminating hypotheses until only the correct one remains

from data is tightly interrelated with specific theoretical commitments about the nature of the biological phenomena under investigation, as well as with experimental practices through which data are produced, tested and modelled. For instance, extracting biologically meaningful inferences from high-throughput genomic data may involve reliance on theories about gene expression and regulation, models of the biological processes being regulated and familiarity with the instruments and organisms from which data were obtained. In this context, "inductive" clearly does not mean "hypothesis-free"; nor can automated reasoning be seen as a substitute to human judgment based on specific expertise and laboratory experience.' (2012, 2)

Certainly, the idea of an entirely theory- or model-free science is absurd. So, Callebaut and Leonelli rightly point out various kinds of theoretical assumptions that enter scientific analyses. But this kind of argument turns out too general and in the end fails to do justice to the remarkable shift towards a strongly inductive approach. Thus, the interesting question is in which ways data-intensive science is indeed theory-laden, and, more importantly, in which sense it can be theory-free. To provide an answer, we now take a detailed look at two algorithms that are widely employed, namely classificatory trees and non-parametric regression. We link these methods to eliminative induction and then determine the kind of theoretical knowledge that has to be presupposed.

4. First case study: classificatory trees

Classificatory trees (e.g. Russell & Norvig 2010, Ch. 18.3.3) are used to determine whether a certain instance belongs to a particular group A depending on a number of parameters C1, …, CN and thus perfectly match the scheme of data-intensive problems as described in Section 2. With help of training data, the tree is set up recursively. First, the parameter CX is determined that contains the largest amount of information with respect to the classification of the training data, as formally measured in terms of Shannon entropy. If CX classifies all instances correctly, the procedure is terminated. Otherwise, two subproblems remain, namely classifying when CX is present and when it is absent. This step is repeated until either all instances are classified correctly or no potential classifiers are left. If the algorithm is successful, the resulting tree structure gives a Boolean expression of necessary and sufficient conditions for A, which can be interpreted as a complex scientific law: e.g. if (C3C2 ∨ C4¬C2)C1 ∨ C6C5¬C1, then A.

The framing of classificatory trees in particular and of problems in data-intensive science in general in terms of a mapping of boundary conditions to an outcome variable fits well with eliminative induction as exemplified in John Stuart Mill's *methods of elimination* (1886, Bk. III, Ch. VIII) with a predecessor in Francis Bacon's *method of exclusion* (1620/1994, Bk. 2). While until the end of the 19th century, Bacon's approach was widely considered the methodological foundation for modern science, eliminative induction has not been very popular since. So, there exist comparably few modern accounts, including von Wright (1951), Mackie (1965, appendix), Skyrms (2000), Baumgartner & Grasshoff (2004), Pietsch (2014).[5]

[5] In the following, I will largely rely on the last account.

In eliminative induction, a phenomenon A is examined under the systematic variation of potentially relevant boundary conditions C1, …, CN with the aim of establishing *causal relevance* or *irrelevance* of these conditions, relative to a certain context or background B consisting of further boundary conditions. The best known and arguably most effective method is the so-called *method of difference* that establishes causal relevance of a boundary condition CX by comparing two instances which differ only in CX and agree in all other circumstances C. If in one instance, both CX and A are present and in the other both CX and A are absent, then CX is causally relevant to A. There is a twin method to the method of difference that one might call the *strict method of agreement*, which establishes causal irrelevance, if the change in CX has no influence on A. Eliminative induction can deal with functional dependencies and an extension of the approach to statistical relationships is straightforward.[6]

Thus, causal (ir-)relevance is a three-place relation: a boundary condition C is (ir-)relevant to a phenomenon A with respect to a certain background B of further conditions that remain constant if causally relevant or are allowed to vary if causally irrelevant. The restriction to a context B is necessary because there is no guarantee that in a different context B*, the causal relation between C and A will continue to hold. Causal laws established by eliminative induction thus have a distinctive contextual or *ceteris-paribus* character. Extensive information about all potentially relevant boundary conditions in as many different situations as possible is necessary to establish reliable causal knowledge by means of eliminative induction. Exactly this kind of information is provided in data-intensive science.

Eliminative induction corresponds to a difference-making account of causality, which is closely related to the counterfactual approach. However, the truth-value of counterfactuals is now determined via the method of difference or the direct method of agreement, and thus by comparison with actual situations that differ from the counterfactual statement only in terms of irrelevant circumstances, and not by a possible-world semantics as in traditional counterfactual approaches like that of David Lewis.

Obviously, classificatory trees rely on eliminative induction. Thus, to assess their quality, one has to look at the premises required for eliminative methods to yield the correct causes. Partial analyses of this problem are given for example in Keynes (1921, Ch. 22), von Wright (1951, Ch. V), Baumgartner & Grasshoff (2004, Sec. IX 2.4), Pietsch (2014, Sec. 3f). We will again follow the exposition in the last reference. There are at least three main assumptions: (i) determinism, i.e. that the phenomenon A is fully determined by boundary conditions C and background B; (ii) constancy of the background, i.e. that no relevant parameters in the background change when two instances are compared via the method of difference or the strict method of agreement; and finally (iii) an adequate vocabulary, that the parameters C reflect suitable causal categories for the given context B. Applied to classificatory trees, we can for example say: if there is a single sufficient condition CX among the C and there is sufficient data in terms of instances of the system in various configurations to avoid spurious correlations, then the classificatory tree algorithm will return CX as cause. Certainly, these

[6] For further discussion, see Pietsch (2014).

assumptions are quite strong. And there are supposedly weaker constraints for causal relations of statistical nature, but this issue goes beyond the scope of the present paper.

We can now identify the elements of theory that have to be presupposed. In particular: (a) one has to know all parameters C that are potentially relevant for the phenomenon A in a given context determined by the background B; (b) one has to assume that for all collected instances and observations the relevant background conditions remain the same, i.e. a stable context B; (c) one has to have good reasons to expect that the parameters C are formulated in stable causal categories that are adequate for a specific research question; (d) there must be a sufficient number of instances to cover all potentially relevant configurations of the phenomenon. If such theoretical knowledge can be established, then there is enough data to avoid spurious correlations and to map the causal structure of the phenomenon without further internal theoretical assumptions about the phenomenon.

This motivates and explains the definition of data-intensive science given in Section 2. In particular, premise (I) is the fundamental condition allowing for a strongly inductive approach based on parameter variation. This viewpoint is further corroborated by the fact that in many cases data-driven approaches become effective rather suddenly—a transition point that could be called a *data threshold* (Halevy et al. 2009). Halevy et al. give a plausible explanation for its existence: 'For many tasks, once we have a billion or so examples, we essentially have a closed set that represents (or at least approximates) what we need, without generative rules.' (2009, 9) At this threshold, the data represents a large fraction of the relevant configurations of the considered phenomenon.

Of course, in scientific practice full theoretical knowledge a) to d) is rarely available. However, in general, including more potentially relevant parameters C will increase the probability that the actual cause of A might be among them, while admittedly also increasing the probability for spurious correlations, i.e. that boundary conditions accidentally produce the right classification. However, more data in terms of instances of different configurations can reduce the probability for such spurious correlations. Thus, more data in terms of parameters and instances will generally increase the chance that correct causal relations are identified by data-intensive algorithms.

5. Second case study: non-parametric regression

A recent paradigm shift in statistics closely mirrors the change from a hypothesis-directed to a more inductive, data-driven approach. It has been described as a transition from *parametric* to *non-parametric* modeling (e.g. Wassermann 2006; Russell & Norvig 2010, Ch. 18.8), from data to algorithmic models (Breiman 2001), or from model-based to model-free approaches. Since the shift concerns methodology and not theoretical or empirical content, it differs in important ways from scientific revolutions. Nevertheless, the statistics community has experienced over the past two decades some of the social ramifications and 'culture clashes' that are typical for scientific paradigm shifts as documented for example in Breiman (2001) or in Norvig's dispute with Noam Chomsky on data-driven machine translation (Norvig 2011).

This paradigm shift has the following basic features: i) Parametric methods usually presuppose considerable modeling assumptions. In particular, they summarize the data in terms of a 'small' number of model parameters specifying for example a Gaussian distribution or linear dependence, hence the name. By contrast, non-parametric modeling presupposes *few modeling assumptions*, e.g. allows for a wide range of functional dependencies or of distribution functions. ii) In non-parametric modeling, predictions are calculated *on the basis of 'all' data*. There is no detour over a parametric model that summarizes the data in terms of a few parameters. iii) While this renders non-parametric modeling quite *flexible* with the ability to quickly react to unexpected data, it also becomes extremely *data- and calculation-intensive*. This aspect accounts for the fact that non-parametric modeling is a relatively recent development in scientific method strongly dependent on advances in information technology. It has largely emerged in parallel with the rise of data-intensive science.

Let me give a simple example as an illustration, the comparison between parametric and non-parametric regression. In a parametric univariate linear regression problem, one has reasonable grounds to suspect that a number of given data points $(x_i;y_i)$ can be summarized in terms of a linear dependency: $y = ax + b$. Thus, two parameters need to be determined, offset b and slope a, which are usually chosen such that the sum of the squared deviations $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$ is minimized.

In non-parametric regression, the data is not summarized in terms of a small number of parameters a and b, but rather all data is kept and used for predictions (e.g. Russell & Norvig 2009, Ch. 18.8.4). A simple non-parametric procedure is *connect-the-dots*. Somewhat more sophisticated is locally weighted regression, in which a regression problem has to be solved for every query point x_q. The y_q-value is determined as $y_q = a_q x_q + b_q$ with the two parameters fixed by minimizing $\sum_{i=1}^{n} K(d(x_q, x_i))(y_i - (a_q x_i + b_q))^2$. Here, K denotes a so-called kernel function that specifies the weight of the different x_i depending on the distance to the query point x_q in terms of a distance function $d()$. Of course, an x_i should be given more weight the closer it is to the query point.

Let us briefly reflect how these regression methods illustrate the differences between parametric and non-parametric modeling i) to iii). While in parametric regression, linear dependency is presupposed as a modeling assumption, the non-parametric method can adapt to arbitrary dependencies. In parametric regression, the nature of the functional relationship has to be independently justified by the theoretical context, which prevents an automation of the modeling process. Certainly, non-parametric regression also makes modeling assumptions, e.g. a suitable kernel function must be chosen that avoids both over- and underfitting. However, within reasonable bounds the kernel function can be chosen by cross-validation. Since often, predictions turn out relatively stable with respect to different choices of kernel functions, an automation of non-parametric modeling remains feasible.

While non-parametric regression is more flexible than parametric regression, it is also much more data-intensive and requires more calculation power. Notably, in the parametric case, a regression problem must be solved only once. Then all predictions can be calculated from the resulting parametric model. In the non-parametric case, a regression problem must be solved for every query point. In principle, each prediction takes recourse to all the data. While the

parametric model consists in a relatively simple mathematical equation, the non-parametric model consists in all the data and an algorithmic procedure for making predictions.

The main difference in terms of theoretical assumptions is that in parametric regression the type of functional dependency is presupposed in contrast to non-parametric regression. The latter again relies on eliminative induction. Essentially, it constitutes a case of Mill's method of concomitant variations, which derives its inferential power from the method of difference as argued for example in Skyrms (2000, Sec. V.9) and Pietsch (2014, Sec. 3d). Thus, the conditions for identifying a causal relationship are largely the same as those discussed in the previous section—determinism, constancy of the background, and correct causal language—resulting in the same premises in terms of theoretical assumptions a)-d). In particular, when mapping a functional dependency, all causally relevant conditions in the background must remain constant. And there must be sufficient data points such that the functional dependence can be traced in adequate detail.

6. Conclusion: data-intensive science and exploratory experimentation

We are finally in a position to evaluate the claims concerning a theory-free science. In both case studies, certain elements of theory had to be presupposed in order to yield reliable results in terms of causal structure that in turn can underwrite successful prediction and manipulation. In particular, among the considered parameters must be those that are causally relevant for a phenomenon in a considered context and not too many that are causally irrelevant to avoid spurious correlations. Also, the parameters should reflect adequate causal categories. Finally, the collected instances or observations should cover all configurations that are relevant in the given context.

Because these aspects all concern the framing of the problem, one could speak of *external* theory-ladenness. By contrast, there is another kind of theory-ladenness that is largely absent from data-intensive science. For example, in classificatory trees no hypotheses are made about the causal connections that link the various parameters. Equally, in non-parametric regression, no assumptions are presupposed about the functional dependencies between different quantities. Thus, the essential difference in comparison with a hypothesis-driven approach is that not much is presupposed about the *internal* causal structure of the phenomenon. Rather, this structure is mapped from the data by parameter variation.

How novel is this approach? On closer scrutiny, data-intensive science much resembles the practice of exploratory as distinguished from hypothesis-directed experimentation (Steinle 1997, Burian 1997, Waters 2007; cp. also Vincenti 1993, 291). Exploratory experimentation essentially consists in the very same parameter variation of eliminative induction, where the experimenter tries to map the system of interest in all those states that she considers relevant. It is this common methodological core, which links exploratory experimentation and data-intensive science and speaks against the claim, for example by Krohs (2012), that the latter constitutes a novel experimental approach focusing on data-gathering.

Not surprisingly, the debate concerning theory-ladenness in exploratory experimentation parallels the discussion in the present article. For example, Steinle (2005) suggests a

distinction between different kinds of theory-ladenness. According to this view, exploratory experimentation presupposes theoretical knowledge in terms of classification systems or empirical rules, but not in terms of theories that postulate empirically inaccessible abstract entities (285). Steinle refers to Duhem, Hacking and Cartwright as having drawn similar distinctions between an experimental/phenomenological and a theoretical level in scientific theories. Indeed, the distinction between exploratory and hypothesis-driven experimentation fits well with Hacking's (1983) claim that experiments have a life of their own and Cartwright's (1983) position of entity realism, which postulates a causal level in science that is mostly phenomenological and largely independent of the theoretical level.

Building on Burian and Steinle's work, Kenneth Waters emphasizes a subtle difference between 'theory-directed' and 'theory-informed'. While in exploratory experimentation, background theories are used 'to set up experiments, generate data, and draw conclusions', such experiments 'are not "directed" by the aim to test, develop, or otherwise articulate an existing theory or hypothesis.' (2007, 280) Laura Franklin makes a similar point that exploratory experiments are theory-laden in terms of background knowledge, but not in terms of local theories (2005, 891).

These remarks closely parallel the previous discussion regarding external and internal theory-ladenness. The distinction between a phenomenological and a theoretical level is also helpful for the analysis of data-intensive science, which supposedly concerns the phenomenological level regarding local, causal structure of phenomena, but does not rise to the theoretical level.

An important difference between exploratory experimentation and data-intensive science is that in the former, data is usually of experimental nature, while the latter often deals with observational data. But this is largely irrelevant from the perspective of a difference-making account of causation according to which experimental intervention has only pragmatic advantages over observational data. Another difference concerns the complexity of the phenomena. While mapping the causal structure by parameter variation is as old as science itself, carrying it out in the computer can address phenomena that were previously largely inaccessible to causal analysis. This new handle, which data-intensive science provides, for mapping the causal structure of highly complex phenomena will make all the difference to scientific practice.

Acknowledgments

I am grateful to Mathias Frisch, Sabina Leonelli, and Sylvester Tremmel for very helpful insights and discussions.

References

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *WIRED Magazine* 16/07.

Bacon, Francis. 1620/1994. *Novum Organum*. Chicago, Il: Open Court.

Baumgartner, Michael & Gerd Graßhoff. 2004. *Kausalität und kausales Schließen*. Norderstedt: Books on Demand.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3):199-231.

Burian, Richard. 1997. "Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938–1952." *History and Philosophy of the Life Sciences* 19:27–45.

Callebaut, Werner. 2012. "Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology." *Studies in History and Philosophy of Biological and Biomedical Science* 43(1):69-80.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Franklin, Laura R. 2005. "Exploratory Experiments." *Philosophy of Science* 72:888-899.

Gray, Jim. 2007. "Jim Gray on eScience: A Transformed Scientific Method." In Tony Hey, Stewart Tansley & Kristin Tolle (eds.). *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Redmond, WA: Microsoft Research.

Hacking, Ian. 1983. *Representing and Intervening.* Cambridge. Cambridge University Press.

Halevy, Alon, Peter Norvig & Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24(2):8-12.

Krohs, Ulrich. 2012. "Convenience Experimentation." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1), 52–57.

Laney, Doug. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety." Research Report. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Leonelli, Sabina. 2012. "Making sense of data-driven research in the biological and biomedical sciences." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43, 1–3.

Mackie, John L. 1980. *The Cement of the Universe.* Oxford: Oxford University Press.

Mill, John S. 1886. *System of Logic*. London: Longmans, Green & Co.

Norvig, Peter. 2011. "On Chomsky and the Two Cultures of Statistical Learning." http://norvig.com/chomsky.html

Norvig, Peter. 2009. "All we want are the facts, ma'am." http://norvig.com/fact-check.html

Pietsch, Wolfgang. 2014. "The Nature of Causal Evidence Based on Eliminative Induction." In P. Illari and F. Russo (eds.), *Topoi*. Doi:10.1007/s11245-013-9190-y

Russell, Stuart & Peter Norvig. 2009. *Artificial Intelligence.* Upper Saddle River, NJ: Pearson.

Skyrms, Brian. 2000. *Choice and Chance*. Belmont, CA: Wadsworth.

Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64:S65–S74.

Steinle, Friedrich. 2005. *Explorative Experimente.* Stuttgart: Franz Steiner Verlag.

Vincenti, Walter. 1993. *What Engineers Know and How They Know It.* Baltimore: Johns Hopkins University Press.

Wasserman, Larry. 2006. *All of Nonparametric Statistics*. New York: Springer.

Waters, C. Kenneth. 2007. "The Nature and Context of Exploratory Experimentation." *History and Philosophy of the Life Sciences* 29(3): 275-284.

von Wright, Georg H. 1951. *A Treatise on Induction and Probability*. New York, NY: Routledge.

Mechanisms and Model-Based fMRI

Mark Povich

Washington University in St. Louis

Abstract. Mechanistic explanations satisfy widely held norms of explanation: the ability to control and answer counterfactual questions about the explanandum. A currently debated issue is whether any non-mechanistic explanations can satisfy these explanatory norms. Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy these norms of explanation. In this paper I will argue that these models are sketches of mechanisms. My argument will make use of model-based fMRI, a novel neuroimaging approach whose significance for current debates on psychological models and mechanistic explanation has yet to be explored.

Word count: 5000

1. Introduction

According to the mechanistic account of explanation, a phenomenon is explained by describing the entities, activities, and organization of the mechanism that produces, underlies, or maintains the phenomenon (see, e.g., Bechtel and Abrahamsen 2005). Mechanistic explanations satisfy what are widely considered normative constraints of explanation: the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon (Craver 2007). These norms capture what is distinctive about the scientific achievement of *explanation* rather than prediction, description, or categorization. A currently debated issue is whether any non-mechanistic forms of explanation can satisfy these explanatory norms.[1] Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy these norms of explanation.

In this paper, in part using recent model-based fMRI research, I will argue that JIM, SUSTAIN, and ALCOVE are in fact mechanism-sketches, i.e. incomplete mechanistic explanations. Model-based approaches to neuroimaging allow cognitive neuroscientists to locate the distributed neural components of psychological models. These novel neuroimaging approaches have developed only recently and philosophers have yet to discuss their significance for current debates on psychological models and mechanistic explanation. The

1 A recent paper arguing affirmatively is Batterman and Rice (2014).

opportunity to demonstrate this significance is one advantage of responding to Weiskopf (2011) in particular.

The paper is organized as follows. In Section 2, I will motivate the mechanistic account of explanation and introduce two crucial concepts in the mechanistic account: the mechanism-sketch and the how-possibly model. In Section 3, I will introduce the models of object recognition and categorization (JIM, SUSTAIN, and ALCOVE) that Weiskopf presents as non-mechanistic, yet explanatory. In Section 4, I will present Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory, and I will begin responding to these arguments. This section demonstrates that JIM is a mechanism-sketch. Demonstrating that SUSTAIN and ALCOVE are mechanism-sketches requires covering recent studies employing model-based fMRI, a novel neuroimaging method that will be explained in section 5.

2. Mechanistic Explanation

Salmon (1984) developed the causal-mechanical account of explanation primarily in response to the covering-law or deductive-nomological model of explanation (Hempel and Oppenheim 1948). According to the deductive-nomological model, an explanation is an argument with descriptions of at least one law of nature and antecedent conditions as premises and a description of the explanandum phenomenon as the conclusion. On this view, explanation is showing that the explanandum phenomenon is predictable given at least one law of nature and certain specific antecedent and boundary conditions. However, tying explanation this closely to prediction generates some famous problems for the covering-law

model (see section 2.3 of Salmon [1989] for a review of these problems). On such a view, many mere correlations come out as explanatory. For example, a falling barometer reliably predicts the weather but the falling barometer does not *explain* the weather. In contrast, on the causal-mechanical view, explanation involves situating the explanandum phenomenon in the causal structure of the world. There are many ways of situating a phenomenon in the causal structure of the world and in this paper I am solely concerned with explanations that identify the mechanism that produces, underlies, or maintains the explanandum phenomenon.[2]

Another problem with tying explanation so closely to prediction is that we miss what is distinctive about the scientific achievement of explanation. Weiskopf (2011) and I agree on what makes explanation distinctive: explanations provide the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon. These are the norms of explanation. Weiskopf and I disagree over what kinds of explanation can satisfy these norms.

Within the mechanistic framework there are two important distinctions: between complete mechanistic models and mechanism-sketches and between how-possibly and how-actually models (Craver 2007). Mechanism-sketches are incomplete descriptions of

2 Other ways of causally situating a phenomenon include etiologically and contextually situating it. See Bechtel (2009) for a discussion of some of these different forms of causal explanation. What Bechtel calls “looking down” I am here calling “mechanistic explanation.”

mechanisms that may contain black boxes and filler terms (Ibid., 113). Mechanistic models rest on a continuum of *more-or-less* complete (114). As more details are incorporated into the model, the more complete it becomes – though no model is ever fully complete, just complete enough for practical purposes. A more complete model is not necessarily a *better* or *more useful* model. There can certainly be *too many* details for the purposes of the modeler and the details that are included should be relevant.[3] Idealization can be readily accommodated within a mechanistic framework.

A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism actually producing, maintaining, or underlying the explanandum phenomenon. As Weiskopf (315) rightly points out, this distinction is epistemic. Turning a how-possibly model into a how-actually model does not require modifying the model itself in any way; it requires testing the model. The greater the evidential support for the model, the more how-actually it is. In contrast, turning a mechanism-sketch into a complete(-enough) model requires modifying the model by filling in missing details.

3. JIM, SUSTAIN, and ALCOVE

In this section I introduce the models of object recognition and categorization JIM, SUSTAIN, and ALCOVE. The next section presents Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory.

3 See Craver (2007, section 4.8) for an account of constitutive (i.e. mechanistic) relevance.

According to JIM (John and Irv's Model), in perception objects are broken down into viewpoint-invariant primitives called "geons". These geons are simple three-dimensional shapes such as cones, bricks, and cylinders. The properties of geons are intended to be non-accidental properties (NAPs), largely unaffected by rotation in depth (Biederman 2000). The geon structure of perceived objects is extracted and stored in memory for later use in comparison and classification.

The importance of NAPs is shown by the fact that sequential matching tasks are extremely easy when stimuli only differ in NAPs. If you are shown a stimulus, then a series of other, rotated stimuli, each of which differs from the first only in NAPs, it is a simple matter to judge which stimuli are the same as or different than the first. Sequential matching tasks with objects that differ in properties that are affected by rotation are much harder.

In JIM, this object recognition process is modeled by a seven layer neural network (Biederman, Cooper, and Fiser 1993). Layer 1 extracts image edges from an input of a line drawing that represents the orientation and depth of an object (182). Layer 2 has three components which represent vertices, axes, and blobs. Layer 3 represents geon attributes such as size, orientation, and aspect ratio. Layers 4 and 5 both derive invariant relations from the extracted geon attributes. Layer 6 receives inputs from layers 3 and 5 and assembles geon features, e.g., "slightly elongated, vertical cone above, perpendicular to and smaller than something" (184). Layer 7 integrates successive outputs from layer 6 and produces an object judgment.

The Attention Learning Covering map (ALCOVE) is a 3-layer, feed-forward, neural network model of object categorization (Kruschke 1992). A perceived stimulus is represented as a point in a multi-dimensional psychological space with each input node representing a single, continuous psychological dimension. For example, a node may represent perceived size, in which case the greater the perceived size of a stimulus, the greater the activation of that node. Each node is modulated by an attentional gate whose strength reflects the relevance of that dimension for the categorization task. Each hidden node represents an exemplar and is activated in proportion to the psychological similarity of the input stimulus to the exemplar. Output nodes represent category responses and are activated by summing hidden nodes and multiplying by the corresponding weights.

The Supervised and Unsupervised Stratified Adaptive Incremental Network (SUSTAIN) is a network model of object categorization similar to ALCOVE (Love, Medin, and Gureckis 2004). Its input nodes also represent a multidimensional psychological space, but they can take continuous and discrete values, including category labels. Like ALCOVE, inputs are modulated by an attentional gate. Unlike ALCOVE, which stores all items individually in memory in exemplar nodes, the next layer of SUSTAIN consists of a set of clusters associated with a category. All of SUSTAIN's clusters compete to respond, with inhibitory connections between each cluster, and the cluster closest to the stimulus in the multidimensional space is the winner. The cluster that wins activates the output unit predicting the category label. The output leads to a decision procedure that generates a category response.

4. Weiskopf's Objections

Weiskopf argues that the previous models are able to satisfy the norms of explanation but are not mechanistic models. How do these models provide the ability to answer counterfactual questions about, and the ability to manipulate and control, the explanandum phenomenon? According to Weiskopf, they satisfy explanatory norms "because these models depict one aspect of the causal structure of the system" (334). This claim is in tension with one reason Weiskopf gives for thinking these models are not mechanistic. He argues, "there may be an underlying mechanistic neural system, but this mechanistic structure is not what cognitive models capture. They capture a level of functional abstraction that this mechanistic structure realizes" (333). But the claim that these models are not mechanistic because they depict a level of functional abstraction, not causal structure, conflicts with the claim that these models are explanatory because they depict causal structure. This conflict results from the general difficulty of specifying how a model can satisfy the norms of explanation without being mechanistic.

One way of trying to reconcile the above claims is to argue that these models are explanatory because they depict causal structure, but they are not mechanistic, because the causal structure that is depicted is not a mechanism. This is the line Weiskopf takes. Why, according to Weiskopf, are these causal structures not mechanisms? He argues that

> If parts [of mechanisms] are allowed to be smeared-out processes or distributed system-level properties, the spatial organization of mechanisms becomes much more difficult to discern. … Weakening the spatial organization constraint by allowing

> distributed, nonlocalized parts incurs costs, in the form of greater difficulty in locating the boundaries of mechanisms and stating their individuation conditions. (334)

The causal structures depicted by JIM, SUSTAIN, and ALCOVE should not be thought of as mechanisms, according to Weiskopf, because these structures are highly distributed. If mechanisms are allowed to contain distributed parts, this will make locating them difficult. The problem, then, is *practical*. Weiskopf does not give any reason to think the *philosophical* (rather than practical) problem of mechanism individuation is made more difficult by allowing distributed parts. [4] Yet numerous neuroimaging methods, especially model-based fMRI, allow cognitive neuroscientists to locate highly distributed neural mechanisms corresponding to the internal variables of computational models. Cognitive neuroscientists are interested in more than the *behavioral* accuracy of these models; they are also interested in their *mechanistic* accuracy. That cognitive neuroscientists conduct neuroimaging studies using these models shows that they are treated as mechanistic. Next I will present some of the neuroimaging studies conducted with JIM and argue that JIM is a mechanism-sketch.

---

4 Weiskopf (331) also cites the phenomenon of neural reuse as inconsistent with mechanism. This assumes that a part of one mechanism cannot be a part of another mechanism but Weiskopf has not provided any reason to think this nor to think that the possibility of reuse should give rise to any special philosophical (rather than practical) problems of mechanism individuation.

JIM was built, not merely to produce the same behavior as human beings in object recognition tasks, but to model something that might really be happening in human brains. Biederman et al. write, "We have concentrated on modeling primal access: The initial activation in a human brain of a basic-level representation of an image from an object exemplar, even a novel one, in the absence of any context that might reduce the set of possible objects" (Biederman, Cooper, Hummel and Fiser 1993, 176). Accordingly, Irving Biederman, one of the co-creators of JIM, and others have conducted various neuroimaging studies to investigate the neural underpinnings of the model.

If JIM is a mechanism-sketch, the systems and processes in the model required for the extraction, storage, and comparison of geon structures must to some extent correspond to (perhaps distributed) components in the actual object recognition mechanisms in the brain. For example, if JIM is a mechanism-sketch, there is an area or a configuration of areas in the brain where simple parts and non-accidental properties are represented. In one study (Hayworth and Biederman 2006), subjects were shown line drawings that were either local feature deleted (LFD), in which every other vertex and line was deleted from each part, removing half the contour, or part deleted (PD) in which half of the parts were removed. On each experimental run, subjects saw either LFD or PD stimuli presented as a sequential pair and had to respond whether or not the exemplars were the same or different. The second stimulus was always mirror-reversed with respect to the first. Each run was comprised of an equal number of three conditions: Identical, Complementary, and Different Exemplar. In the Identical condition, the second stimulus was the same as the first stimulus (mirror-reversed,

as all of the second stimuli were). In the Complementary condition, the second stimulus was the complement of the first, where an LFD-complement is composed of the deleted contour of the first and a PD-complement is composed of the deleted parts of the first. In the Different Exemplar condition, the second stimulus is a line-drawing of a different exemplar than the first.

An fMRI-adaptation design was used, which "relies on the assumption that neural adaptation reduces activity when two successive stimuli activate the same subpopulation but not when they stimulate different subpopulations" (Krekelberg, Boynton, van Wezel 2006, 250; see also Kourtzi and Grill-Spector 2005). The results of the study showed adaptation between LFD complements and lack of adaptation between PD complements in lateral occipital complex, especially the posterior fusiform area, an area known to be involved in object recognition. These results imply that this area is "representing the parts of an object, rather than local features, templates, or object concepts" (Hayworth and Biederman 2006, 4029). Biederman has conducted many other fMRI experiments, including some that "suggest that LO [lateral occipital cortex] is the locus of the neural correlate for the greater detectability for nonaccidental relations" (Kim and Biederman 1824).

While these results resolve Weiskopf's worry about the difficulty of locating distributed parts, he has another argument for why JIM is not mechanistic. JIM has properties that do not and could not correspond to anything in the brain. Weiskopf (2011, 331) mentions JIM's "Fast Enabling Links" (FELs), which allow the model to bind representations and which have infinite propagation speed. According to Weiskopf, FELs are an example of

fictionalization, "putting components into a model that are known not to correspond to any element of the modeled system, but which serve an essential role in getting the model to operate correctly" (Ibid.), and he argues that this undermines the claim that JIM is a mechanism-sketch. Weiskopf is right that FELs are an essential fictionalization, but playing an essential role in getting a model to operate is not the same as explaining; these parts of the model carry no explanatory information and render the model, or at least part of it, how-possibly (where the possibility involved is not physical possibility, since FELs are physically impossible). Right now FELs play the black box role of whatever-it-is-that-accounts-for-binding. In addition to playing a black box role, they serve practical and epistemic purposes like the ones discussed by Bogen (2005), such as suggesting, constraining, and sharpening questions about mechanisms. Let me explain how by comparing FELs to Bogen's example of the GHK equations.

The Goldman, Hodgkin, and Katz (GHK) voltage and current equations are used to determine the reversal potential across a cell's membrane and the current across the membrane carried by an ion. These equations rely on the incorrect assumptions that each ion channel is homogeneous and that interactions among ions do not influence their rate (Bogen 409). About the inadequacy of these equations Bogen writes,

> While some generalizations are useful because they deliver empirically acceptable quantitative approximations, others are useful because they do not… Investigators used these and other GHK equation failures as problems to be solved by finding out more about how ion channels work. Fine-grained descriptions of exceptions to the

> GHK equations and the conditions under which they occur sharpened the problems and provided hints about how to approach them. (Bogen 410)

The GHK equations provide a case of "using incorrect generalizations to articulate and develop mechanistic explanations" (Bogen 409). I argue that something similar can be said about FELs. Not only do FELs play an essential black box role, FELs suggest new questions about mechanisms, new problems to be solved. For example, Hummel and Biederman (1992) write,

> [T]he independence of FELs and standard excitatory-inhibitory connections in JIM has important computational consequences. Specifically, this independence allows JIM to treat the constraints on feature linking (by synchrony) separately from the constraints on property inference (by excitation and inhibition). That is, cells can phase lock without influencing one another's level of activity and vice versa. Although it remains an open question whether a neuroanatomical analog of FELs will be found to exist, we suggest that the distinction between feature linking and property inference is likely to remain an important one. (510)

Like the GHK equations, FELs suggest new lines of investigation, in this case regarding the relation between feature linking, property inference, and their neural mechanisms. Specifically, FELs suggest questions such as, "Can biological neurons phase lock without influencing one another's activity?" and "Are there other ways biological neurons could implement feature linking and property inference independently?".

In the next section, I will explain model-based fMRI and demonstrate how recent model-based fMRI studies show that SUSTAIN and ALCOVE are mechanism-sketches.

5. Model-Based fMRI

Model-based fMRI is a neuroimaging method that aims to discover the neural mechanisms that correspond to model variables. Model-based fMRI "can be used as a means of discriminating between competing computational models of cognitive and neural function. Thus, model-based fMRI provides insight into 'how' a particular cognitive function might be implemented in the brain, not only 'where' it is implemented" (O' Doherty, Hampton, and Kim 39). In this way, model-based fMRI provides a way of discriminating between competing, equally behaviorally confirmed cognitive models (Glascher and O'Doherty 502).

Functional magnetic resonance imaging (fMRI) is a neuroimaging method that provides an indirect measure of neuronal activity. Neuronal activity requires glucose and oxygen for fuel, which the vascular system provides. The oxygen is bound to hemoglobin molecules and the magnetic properties of deoxygenated hemoglobin are detectable by fMRI. In this way, fMRI measures a physiological indicator of oxygen consumption – deoxyhemoglobin concentration – that correlates with changes in neuronal activity (Huettel, Song, and McCarthy 159-160).

To conduct a model-based fMRI analysis, one starts with a computational model that describes the function(s) by which stimuli are transformed to result in behavioral output. Stimulus input and behavioral output are observable, but the computational model postulates internal variables linking input and output. The neural correlates of these internal variables, at

each time point in the experiment, can then be located using regression analyses (O' Doherty, Hampton, and Kim 36).

The variables that change from trial to trial are converted into a time series of the model-predicted BOLD (blood-oxygen-level dependent) response and then convolved with a canonical hemodynamic response function (Glascher and O'Doherty 505). This just means that the predicted variable values, taken over time, are mathematically combined with a stereotypical BOLD signal function. This is done to account for the usual lag in the hemodynamic response (O' Doherty, Hampton, and Kim 37). This yields a new function that, when put into a general linear model, can be regressed against fMRI data. General linear models have the following form:

$$y = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n + e$$

where y is the observed data, the x_i are regressors (the model-predicted time series), the B_i are variable weights (B_0 represents the contribution of factors held constant throughout the experiment), and e is residual noise in the data (Huettel, Song, and McCarthy 343). This allows researchers to identify brain areas where the model-predicted time series significantly correlates with the observed BOLD signal changes over time.

I should make clear that model-based fMRI has limitations and does not obviate the need for other neuroimaging methods (e.g., PET, EEG, or MEG). Like fMRI in general, model-based fMRI can only establish correlations between neural activity and behavior. In order to establish causal claims about neural activity and behavior, the same methods need to be used that were used before the introduction of model-based fMRI, such as lesioning and

transcranial magnetic stimulation (TMS) (O' Doherty, Hampton, and Kim 50). Like fMRI in general, model-based fMRI also has poor spatiotemporal resolution. This means that small computational signals such as those at the level of the single neuron will go undetected by model-based fMRI. For these reasons, a model-based approach to other neuroimaging methods is needed (Ibid.)

Now that we have a basic understanding of how model-based fMRI works and what it can accomplish, let me return to SUSTAIN and ALCOVE and show how they are mechanism-sketches by drawing on recent model-based fMRI research.

Both models were investigated in a model-based fMRI study in which participants completed a rule-plus-exception category learning task (Davis, Love, and Preston 2012). During the task, a schematic beetle was presented and subjects were asked to classify it as "Hole A" or "Hole B," after which they received feedback. The beetles varied on four of the following five attributes, with the fifth held constant: eyes (green or red), tail (oval or triangular), legs (thin or thick), antennae (spindly or fuzzy), and fangs (pointy or round). Six of the eight beetles presented could be correctly categorized on the basis of a single attribute. For example, three out of four Hole A beetles might have thick legs and three out of four Hole B beetles could have thin legs. The other beetles were exceptions to the rule, having legs that appeared to match the other category.

Two predictions from SUSTAIN and ALCOVE were tested. First, during stimulus presentation SUSTAIN predicts a recognition advantage for exceptions but ALCOVE predicts no recognition advantage. This is called the recognition strength measure. This

difference in recognition strength measure predictions arises because in ALCOVE, but not in SUSTAIN, all items are stored individually in memory regardless of whether they are exceptions or rule-following items. Second, when subjects are given feedback, both SUSTAIN and ALCOVE predict that exceptions should lead to greater prediction error. This is called the error correction measure (Ibid., 263-4).

The results showed that the recognition strength measures and error correction measures predicted by SUSTAIN found correlations in MTL regions including bilateral hippocampus, parahippocampal cortex, and perirhinal cortex, and regions in bilateral hippocampus and perirhinal cortex, respectively. ALCOVE's predicted recognition strength measures did not find correlations in MTL, although its error correction predictions found correlations in MTL similar to SUSTAIN's (Ibid., 266-7). These results "suggest that, like SUSTAIN, the MTL contributes to category learning by forming specialized category representations appropriate for the learning context" (Davis, Love, and Preston 269). Furthermore, these correspondences to brain areas open a whole new range of opportunities for manipulation and provide answers to counterfactual questions that were not available before, thereby increasing the explanatory power of these models.

SUSTAIN and ALCOVE are mechanism-sketches. SUSTAIN is more how-actually than ALCOVE because both of SUSTAIN's prediction measures (recognition strength and error correction) were significantly correlated to areas of brain activation, whereas only one of ALCOVE's (error correction) was correlated. SUSTAIN, therefore, has more evidential support than ALCOVE. These results also show that cognitive neuroscientists are currently

advancing the ability to map the entities and activities in psychological models to distributed neural systems, such as MTL regions spanning bilateral hippocampus, parahippocampal cortex, and perirhinal cortex.

Davis, Love, and Preston (2012) are at times quite explicit about the mechanistic nature of the models they are investigating, although they do not use the term "mechanistic." For instance, they write, "We use a model-based functional magnetic resonance imaging (fMRI) approach to test the proposed mapping between MTL function and SUSTAIN's representational properties" (261) and "The theory we forward relating SUSTAIN to the MTL…goes beyond the model's equations by tying model operations to brain regions" (270). Given their emphasis on mapping models to the brain, it is clear that they intend the models to be mechanistic. They are interested in more than the behavioral accuracy of these models. SUSTAIN and ALCOVE are already behaviorally well-confirmed, but model-based fMRI allowed Davis et al. to test their mechanistic accuracy**.**

6. Conclusion

Weiskopf (2011) presented three models of object recognition and categorization, JIM, ALCOVE, and SUSTAIN, that he claimed were non-mechanistic, yet explanatory. He argued that they were not mechanistic because their parts could not be neatly localized and they contained some components, such as Fast Enabling Links (FELs), which could not correspond to anything in the brain but are nevertheless essential for the proper working of the model. I argued on the contrary that these models are mechanism-sketches. In addition to

playing a black box role, FELs possess non-explanatory virtues such as suggesting new lines of investigation about feature linking and property inference.

My argument for the claim that SUSTAIN and ALCOVE are mechanism-sketches relied on model-based fMRI research. Model-based fMRI and other model-based neuroimaging approaches are beginning to allow cognitive neuroscientists to map psychological models onto the brain. Cognitive neuroscientists can then discriminate between equally behaviorally confirmed psychological models. The development of these model-based approaches has broader implications, beyond the narrow dispute over JIM, SUSTAIN, and ALCOVE, for the debate over the explanatory and mechanistic status of psychological models. As cognitive neuroscientists continue to test psychological models against neuroimaging data using model-based techniques, they will retain those models that find correspondences in the brain and reject those that do not, and in so doing reveal that explanatory progress in cognitive neuroscience consists in the development of increasingly mechanistic models.

References

Batterman, Robert and Collin Rice. 2014. "Minimal Model Explanations." *Philosophy of Science* 81.3: 349-376.

Bechtel, William. 2009. "Looking Down, Around, and Up: Mechanistic Explanation in Psychology." *Philosophical Psychology* 22.5: 543-64.

Bechtel, William and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of the Biological and Biomedical Sciences* 36.2: 421-41.

Biederman, Irving. 2000. "Recognizing Depth-rotated Objects: A Review of Recent Research and Theory." *Spatial Vision* 13.2,3: 241-53.

Biederman, Irving, Eric E. Cooper, John E. Hummel, and Jozsef Fiser. 1993. "Geon Theory as an Account of Shape Recognition in Mind, Brain and Machine." In *Proceedings of the 4th British Machine Vision Conference*, ed. John Illingworth, 175-86. London: Springer-Verlag.

Bogen, Jim. 2005. "Regularities and Causality; Generalizations and Causal Explanations." *Studies in History and Philosophy of Biology and Biomedical Sciences* 36: 397-420.

Craver, Carl. 2007. *Explaining the Brain.* Oxford: Oxford University Press.

Davis, Tyler, Bradley C. Love, and Alison R. Preston. 2012. "Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members." *Cerebral Cortex* 22: 260-73.

Glascher, Jan P. and John P. O' Doherty. 2010. "Model-based Approaches to Neuroimaging: Combining Reinforcement Learning Theory with fMRI Data." *WIREs Cognitive Science* 1: 501-10.

Hayworth, Kenneth J. and Irving Biederman. 2006. "Neural Evidence for Intermediate Representations in Object Recognition." *Vision Research* 46: 4024-31.

Hempel, Carl G. and Paul Oppenheim. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15: 135-75.

Huettel, Scott A., Allen W. Song, and Gregory McCarthy. 2009. *Functional Magnetic Resonance Imaging*. Sunderland, Mass.: Sinauer Associates.

Hummel, John E., and Irving Biederman. 1992. "Dynamic Binding in a Neural Network for Shape Recognition." *Psychological Review* 99: 480-517.

Kim, Jiye G. and Irving Biederman. 2012. "Greater sensitivity to nonaccidental than metric changes in the relations between simple shapes in the lateral occipital cortex." *NeuroImage* 63: 1818-1826.

Kourtzi, Zoe and Kalanit Grill-Spector. 2005. "fMRI Adaptation: A Tool for Studying Visual Representations in the Primate Brain." In *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision,* ed. Colin W. G. Clifford and Gillian Rhodes, 173-88. New York: Oxford University Press.

Krekelberg, Bart, Geoffrey M. Boynton and Richard J.A. van Wezel. 2006. "Adaptation: From Single Cells to BOLD Signals." *TRENDS in Neurosciences* 29.5: 250-56.

Kruschke, John K. 1992. "ALCOVE: An Exemplar-based Connectionist Model of Category Learning." *Psychological Review* 99: 22-44.

Love, Bradley C., Douglas L. Medin, and Todd M. Gureckis. 2004. "SUSTAIN: A Network Model of Category Learning." *Psychological Review* 111: 309-32.

Love, Bradley C. and Todd M. Gureckis. 2007. "Models in Search of a Brain."

O' Doherty, John P., Alan Hampton, and Hackjin Kim. 2007. "Model-Based fMRI and Its Application to Reward Learning and Decision Making." *Annals of the New York Academy of Sciences* 1104: 35-53.

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, Wesley C. 1989. "Four Decades of Scientific Explanation." In *Minnesota Studies in the Philosophy of Science, Vol 13: Scientific Explanation*, ed. Wesley Salmon and Philip Kitcher, 3-219. Minneapolis: University of Minnesota Press.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183.3: 313-38.

Philosophy of Science

Atrazine Research and Criteria of Characterizational Adequacy

--Manuscript Draft--

Manuscript Number:	12372
Full Title:	Atrazine Research and Criteria of Characterizational Adequacy
Article Type:	PSA 2014 Contributed Paper
Keywords:	Atrazine; Toxicology; Endocrine Disruption; Reductionism; Evo-Devo; Evolutionary Biology
Corresponding Author:	John Powers UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	John Powers
First Author Secondary Information:	
Order of Authors:	John Powers
Order of Authors Secondary Information:	
Abstract:	Abstract: The effects of atrazine on amphibians has been the subject of much research, requiring the input of many disciplines. Theory reductive accounts of the relationships among scientific disciplines do not seem to characterize well the ways that diverse disciplines interact in the context of addressing such complex scientific problems. "Problem agenda" accounts of localized scientific integrations seem to fare better. However, problem agenda accounts have tended to focus rather narrowly on scientific explanation. Attention to the details of atrazine research reveals that characterization deserves the sort of attention that problem agenda theorists have thus far reserved for explanation.

Manuscript

Atrazine Research and Criteria of Characterizational Adequacy

Abstract: The effects of atrazine on amphibians has been the subject of much research, requiring the input of many disciplines. Theory reductive accounts of the relationships among scientific disciplines do not seem to characterize well the ways that diverse disciplines interact in the context of addressing such complex scientific problems. "Problem agenda" accounts of localized scientific integrations seem to fare better. However, problem agenda accounts have tended to focus rather narrowly on scientific explanation. Attention to the details of atrazine research reveals that characterization deserves the sort of attention that problem agenda theorists have thus far reserved for explanation.

Word Count: 4994

1. Background and introduction

Although a consensus has developed around skepticism about the prospects and motivation for Nagelian theory reduction in the biological sciences, several authors have pointed out that participants in this consensus have historically failed to offer much in the way of well-developed alternative philosophical accounts of how various sciences and disciplines might be epistemically related (Rosenberg 1997, Robert 2004); in response to the apparent untenability of theory reduction, proposals of the epistemic relationships among the various biological and allied disciplines have typically been given in terms of explanatory reductionist, anti-reductionist, and nonreductionist (often pluralist) strategies, but a need persists for detailed development of these strategies and application to particular case studies (Brigandt and Love 2012). Contra more radically permissive pluralist accounts (e.g., Dupre 1993), advocates of the so-called "pluralist stance" have contended that the nature of the specific scientific problem or question being addressed constrains the "variety of acceptable classificatory or explanatory schemes." (Kellert et al 2006) Taking onboard this feature of the pluralist stance, Love (2008) and Brigandt (2010) have offered structured accounts of local integrations in evolutionary developmental biology (evo-devo) that are centered around solving particular problems and explaining particular explananda. These local integrations need not, for the authors, necessarily be part of any broader unificatory theoretical reduction

of the sort envisioned by proponents of theory reduction (Nagel 1961; Schaffner 1993) or unificatory explanatory ideal (Kitcher 2001). Love and Brigandt's views do, however, emphasis the important role of more problem-specific explanatory (as opposed to theoretical) reductions in biological explanation. Where theory reduction approaches tend to contend that laws describing "lower" mereological levels are always more fundamental in explanation, on the problem-centered view, explanatory fundamentality "*varies with the specific problem* at hand." (Brigandt 2010) Thus, Brigandt and Love's problem-centered integrative frameworks are *nonreductionist* in that they do not necessarily ascribe explanatory fundamentality to lower level epistemic units (laws, theories, models, etc.). However, these frameworks are not *antireductionist* because they reserve a place for reductive explanation when such explanation is called for by the nature of the specific scientific problem or problems under consideration.

Love and Brigandt both take research into *explanations* of evolutionary innovation and novelty as their focus. Hence, Love's and Brigandt's accounts of local integration have centered on questions about multidisciplinary explanation (Love 2008; Brigandt 2010). But while explanation is the *central* concern of many biological projects, explanation is not the *only* concern. Waters (2007) points out that the findings of so-called "exploratory" experiments can have significance for various scientific goals other than explanation and theory development, including knowledge about experimental manipulation and conceptual development to guide future research. Minimally, explanation requires explananda, and those explananda often require scientific investigations to in order to be recognized as things wanting

explanation and to disclose the ways in which they might be experimentally manipulated or exploited in the future. Thus, solving scientific problems often requires a certain sort of characterization, achieved through various scientific practices, that is conceptually distinct from explanation. Maps of concentrations of environmental pollutants, for instance, are an outcome of scientific experiments and modeling practices where the goal is experimentally-grounded pattern characterization rather than the provision of an explanatory account of a phenomenon, although the pattern so characterized may later be an object of explanation.

Love (2008) and Brigandt (2010) do not explicitly treat the sort of empirical characterization that I've characterized above (although their projects appear amenable to the inclusion of such a treatment). Love 2008's nonreductionist "problem agenda" account of local (as opposed to more broadly theoretical or unificatory) integration in the biological sciences deploys the concept of "criteria of explanatory adequacy." These criteria, associated with particular problems and sets of problems, act as unifying constraints by specifying what sorts scientific explanations are adequate for the problems that motivate them. I here seek to augment nonreductionist problem-centered epistemologies of multidisciplinary integration with a treatment of the criteria by which various scientific disciplines might judge empirical characterization*s* (as opposed to explanations) and the processes by which such characterizations are generated to be adequate in the context of solving particular problems and sets of problems.

Research into the endocrine disrupting effects of the herbicide, atrazine, is a promising case study because, while concerned with explanation, this research agenda clearly also makes necessary use of empirical characterization (e.g., dose-response curves for atrazine exposure and geographic maps of atrazine concentration). Additionally, atrazine research is remarkably multidisciplinary. Critical evaluation of claims about inherent multidisciplinarity on the part of scientists participating in atrazine research provides an opportunity for describing how certain disciplines play a role in offering answers to the questions that atrazine researchers seek to answer. Articulating the roles played by the contributions of each discipline also presents an opportunity to demonstrate how a nonreductionist epistemology can provide an account of disciplinary integration centered on solving particular problems and answering particular questions. Such an account is desirable not only because it promises to fill the void left by the abandonment of traditional theory reduction approaches for describing epistemic relationships among disciplines, but also because it promises to yield novel insights into reasoning across scientific disciplines and novel interpretations of multidisciplinary disagreement.

2. Atrazine research as a case study

Atrazine is a top selling herbicide that is a persistent and widely distributed ground and surface water pollutant. The effects of atrazine on amphibians and the contribution of these effects to global amphibian decline has been the subject of much research, requiring the input of many disciplines. Work in molecular biology, biochemistry, developmental biology, endocrinology, physiology, and organismal

biology has revealed that atrazine acts as an endocrine (hormone) disruptor in vertebrate organisms; it induces a class of enzymes (aromatases) that convert androgens (e.g., testosterone) into estrogens (e.g. estradiol). This conversion has diverse effects on different kinds of vertebrate organisms, from "feminization" leading to decreased reproductive success in frogs to increased cancer rates in humans. Describing and predicting atrazine persistence, transport, and exposure has involved input from diverse disciplines including hydrology, agricultural science, geology, soil science, environmental chemistry, and meteorology (Hayes 2005, Hayes *et al* 2011). Tyrone Hayes, a leading researcher on atrazine's endocrine disrupting effects on frogs, claims that,

> "To truly assess the impact of atrazine on amphibians in the wild, diverse fields of study including endocrinology, developmental biology, molecular biology, cellular biology, ecology, and evolutionary biology need to be invoked. To understand fully the long-term impacts on the environment, meteorology, geology, hydrology, chemistry, statistics, mathematics and other disciplines well outside of biology are required." (2005, 321)

Although understanding physiological developmental mechanisms seems key to understanding abnormal amphibian development resulting from exposure to endocrine disruptors like atrazine, and research on atrazine transport and persistence seems clearly necessary to infer exposure rates and magnitudes, it is not immediately clear what it is about this question that requires input from other disciplines, e.g. evolutionary biology. What justifies Hayes' claim that evolutionary biology is *required*? A framework for structuring multidisciplinary inputs within the atrazine research program can help us articulate the roles played by various disciplines in answering the question of the impact of atrazine's endocrine

disrupting effects on amphibians in the wild and thereby allow us to critically evaluate claims (like Hayes's) for the necessity of particular disciplines.

Love (2008) develops an account of localized integration in the sciences based on what he calls "problem agendas," or sets of problems (complex questions composed of simpler questions) related to a particular epistemic goal. Here I cast the impact of atrazine's endocrine disrupting effects on amphibians as a simpler question within the problem (complex question) of the impact of anthropogenic endocrine disruptors on the environment. I will characterize environmental endocrine disruption as a problem shared by the problem agendas of environmental toxicity and developmental endocrine function. To aid in this characterization, I will describe Love's notion of "explanatory adequacy," the criteria by which explanatory answers to problems (complex questions) on a particular problem agenda are judged to be adequate or inadequate. I will then introduce the complementary concept of *criteria of characterizational adequacy* (CCA), criteria by which empirically grounded characterizations and the practices by which they are generated are judged to be adequate or inadequate with respect to particular epistemic goals. I will show how the criteria of characterizational and explanatory adequacy of the two problem agendas of environmental toxicity and developmental endocrine function structure disciplinary inputs with respect to the narrower question of the impact of atrazine's endocrine disrupting effects on amphibians. Finally, I will show how a set of proposed criteria of explanatory and characterizational adequacy drawn from the two problem agendas can make clearer

the contributions of evolutionary biology to the question of the impacts of atrazine on amphibians in the wild.

3. Love on local integration

Love (2008) characterizes *problem agendas* as sets of problems (complex questions) related to a particular complex epistemic goal. Problem agendas are united in part by *criteria of explanatory adequacy*, criteria for judging the acceptability of candidate solutions to the problems composing the agenda (875). Against theorists who argue for (typically reductive) stable theoretical integration or unification of diverse fields of science (Nagel 1961; Schaffner 1993), Love argues that integration of multiple fields of study can profitably be localized to particular epistemic goals without necessarily requiring more global theoretical integration or unification.

Criteria of explanatory adequacy are central to Love's account of localized integration. Such criteria make possible "an explicit account of how different areas of research make their contribution without one being more fundamental than another." (2008, 875) Because calls for multidisciplinary research typically arise out of the need to solve problems and answer questions rather than a need for theory-building or testing, what is needed is an account of what ought to count as adequate answers to the complex questions driving the research.

Love uses the problem agenda of evolutionary innovation and novelty as an example to illustrate the concepts of problem agendas and criteria of explanatory adequacy. Problems on the innovation and novelty agenda include, e.g., "How did vertebrate jaws originate?" and "How did avian flight originate?" Although perhaps

superficially resembling more ordinary questions (e.g., How was the window broken?) these problems are not standard interrogatives of the sort that can be answered with a single proposition. These problems, due to their complexity and the diversity of simpler questions that they naturally engender, are thought to require multidisciplinary input from developmental, evolutionary, molecular, and systematic biology (2008, 879).

Love claims that the inputs of these disciplines can be structured by the criteria of explanatory adequacy associated with the project. For Love, adequate explanations of the origination of radical evolutionary changes in phenotype must meet three criteria grounded in the nature of the explananda. First, the explanation must address both form and function; e.g., explanations of the origination of vertebrate jaws must include considerations related to how these sorts of jaws function given the particular forms that they take. Second, accounts of origination must explain innovation and novelty at all biological levels of organization as well as relations among these levels, e.g., genetic, cellular, modular, organismal, and population levels (Love 2008, 880). And finally, there is the third criterion of "degree of generalization," which deals with how different problems within the agenda are related. For the case of evolutionary novelty, this criterion can be broken into two further questions. 1- "Can investigations of particular novelties be generalized to other research on different innovations or novelties?" and 2- "Can investigations of model systems be generalized to the phylogenetic juncture relevant to the innovation or novelty under scrutiny?" (Love 2008, 881) The

concern here is the appropriateness of generalizations from one problem or question within the agenda to others.

4. The problem of endocrine disruptors in the environment

Environmental problems are exemplary of the sorts of problems that require multi-disciplinary input for generating adequate solutions (Love 2008, 875). The question of atrazine's effects on amphibians in the wild as a result of its endocrine disrupting properties can be viewed as a simpler question located within the environmental problem (complex question) of endocrine disruptors and their ecological impacts. This problem is shared by the problem agendas of environmental toxicity and developmental endocrine function, each with its own criteria of explanatory and characterizational adequacy. These criteria will be shown to constrain and unify attempts at answering questions clustered around the impact of atrazine on amphibians. To illustrate this, I will begin by offering some plausible sample questions germane to the broader question of atrazine's role as an environmental amphibian endocrine disruptor. Notice that the levels of biological organization at which the questions are aimed increases sequentially. The first question is aimed at the biochemical and genetic levels; the second is aimed at the morphological level; the third is aimed at the population level, and the fourth is aimed at global scale ecological phenomena and impacts on higher-level taxa. The species named in the first through the third question are reflective of some of the organisms that are frequently used in such research (Hayes 2005; 2011).

1. What effect does atrazine exposure at a given concentration and duration have on CYP19 (aromatase gene) expression in *Xenopus laevis*?

2. How do the morphological effects of given concentration and duration of atrazine exposure in *Hyperolius argus* differ depending on the developmental stage at which exposure occurs?
3. What impacts do atrazine's endocrine disrupting effect have on *Rana pipiens* populations in Midwestern corn growing regions?
4. Does atrazine's endocrine disrupting effect play a significant role in global amphibian decline?

I want to suggest that answers to these and similar questions will be constrained by criteria of explanatory and characterizational adequacy drawn from the two problem agendas in which problem of environmental endocrine disruption and the question of atrazine's impact on amphibians seem to reside.

5. Environmental toxicity (and a distinction between explanation and empirical characterization)

Environmental toxicology has been described as "the study of the impacts of pollutants on the structure and function of ecological systems." (Landis *et al.* 2010, 1) Its focus is the identification of toxic agents and the establishment of the causal bases of their toxicity (Landis *et al.* 2010, Chapter 3).

These two epistemic goals highlight a distinction between empirical characterization and explanation. In the case of identifying toxic agents, the goal is identifying and characterizing the effects of a chemical and classifying it according to its toxic properties, a task of description and evaluation (characterization). In the case of identifying causal bases of toxicity, the goal is explanatory, concerned with providing a causal account of the processes by which a chemical gives rise to toxic

effects. Such explanatory goals seem clearly amenable to constraint by criteria of explanatory adequacy as Love develops the concept. E.g., an explanation of the mechanism by which atrazine is toxic to plants, starvation and harmful oxidative effects due to interruption of plastoquinone-binding in photosystem II (Appleby *et al*. 2001), is constrained by the criterion of explaining higher-level physiological effects by reference to lower-level biochemical processes.

It seems strange, however, to say that the descriptive and evaluative goals of describing and classifying chemicals and their impacts according to toxicity are constrained by *sensu stricto* criteria of *explanatory* adequacy. After all, the goal is description and classification rather than explanation (although as we will see, some descriptive and classificatory claims derive their inferential justification from explanatory accounts). Rather, such attempts at scientific characterization are constrained by what we might call *criteria of characterizational adequacy.* Criteria of characterizational adequacy (CCA) are constraints on empirically grounded characterizations (e.g. claims about response, correlation, concentration, *etc*.) that specify what counts as adequate justification for those sorts of characterizations.

To illustrate how the concept of CCA might apply, consider the case of dose-response curves common to the problem agenda of environmental toxicology. A dose response curve is "a graph describing the response of an enzyme, organism, population, or biological community to a range of concentrations of a xenobiotic." (Landis *et al.* 2010, 36) The task here is characterizational rather than explanatory; such curves have no necessary reference to causal mechanisms explaining the phenomena represented by the graph. However, the production of such a

characterization is constrained by certain criteria. For example, the points on the graph must make reference to a concentration of the xenobiotic and must be compared to a control in which the xenobiotic is absent, i.e., the "normal" behavior of the enzyme, organism, population, or biological community under consideration. These "concentration relative" and "compared to control" CCA allow us to see the contributions of exploratory research aimed at characterizing the properties of entities in a way that we could not if we considered only criteria of explanatory adequacy. Much of the research activity in the environmental toxicity problem agenda is aimed at characterizing concentrations (in cells, organs, organisms, particular habitats, *etc.*)(Rohr and McCoy 2010, Hayes *et al*. 2011). Properties of entities at characterized concentrations must then be compared to properties of entities free from the putative toxin, and the characterization of these toxin-free properties involves exploratory research. In the case of atrazine, the near ubiquity of the chemical in fresh water supplies, and its potential for effects at very low doses has necessitated the development of sophisticated filtering techniques and careful attention to laboratory hygiene in order to characterize the properties of biological entities in their atrazine-free conditions. Additionally, due again to atrazine's near ubiquity in the environment, the "compared to control" criterion has made essential early characterizations of frog morphology in the wild (e.g. Witschi 1929), characterizations made before the wide-spread application of atrazine began in the 1950s (Hayes 2004; Rohr and McCoy 2010).

6. Developmental endocrine function

The purpose of the study of developmental endocrine function is to provide an account of the biochemical processes and pathways of hormone synthesis, storage, and physiological function during organismal development (Hayes 2005). Problems (complex questions) comprising a developmental endocrine function problem agenda include "how do sex steroids control development?" and "how do thyroid hormones control development?" Solutions to these sorts of problems would seem to be constrained by the need to address causality at multiple levels of biochemical and biological organization and the need to justify generalizations from insights about pathways and processes in model organisms to claims about other organisms (roughly, Love's second and third criteria of explanatory adequacy) (Love 2008, 880-881).

Research into sex steroid determination of sexual development provides examples of these criteria in action. Comparative endocrinology research has discovered that androgens and estrogens control sexual development across the vertebrates, although the developmental effects of these hormones vary by taxa, imposing limits on generalizations made across taxa. The effects of these hormones tend to be "organizational" and irreversible at earlier stages of development and "activational" and reversible in adults. Explanations of these effects (and their relative permanence) make reference to biochemical pathways, gene expression, cellular metabolism and differentiation, and organ development (Hayes 2005, Hayes *et al* 2011).

7. Criteria of adequacy applicable the problem of endocrine disruptors in the environment and the narrower atrazine question

Now I wish to show how some of the criteria of explanatory and characterizational adequacy that constrain solutions to problems on the environmental toxicology and developmental endocrinology problem agendas also constrain answering narrower questions germane to assessing atrazine impacts on amphibians. First, because environmental toxicity problem solutions must make reference to controls free from the putative toxin, answers to the question of the impacts of atrazine must be predicated on atrazine exposure effects compared to atrazine-free controls or hypothetical populations. Much of the important research in the "emerging" science of amphibian endocrine disruption has been made possible by basic research on, e.g., CYP19 gene expression, aromatase catalysation of estrogenesis, sex steroid control of sexual differentiation during amphibian development, amphibian reproductive anatomy and behavior, and population genetic modeling of amphibian evolution (2005, Hayes *et al* 2011) Such studies provide a *baseline characterization* against which the effects of atrazine at environmental concentrations inferred by sampling (as well as transport and persistence studies) can be compared. This criterion also provides grounds for the rejection of some proposed answers to questions about atrazine's effects on amphibians. Some authors, for instance, have proposed that hermaphrodism is widespread in wild amphibian populations in the absence of atrazine exposure (Carr and Solomon 2003). However, this conclusion was based on field and laboratory studies in which the controls are thought to have been exposed to environmental atrazine, possible at relatively high concentrations (Hayes 2004; 2005, Rohr and McCoy 2008)

Second, adequate answers to questions about atrazine's effects on amphibians must give an account of all the relevant levels of biological organization. For instance, an answer to the third question in the list above would plausibly give a causal account of the effects of atrazine on Midwestern leopard frog populations by invoking atrazine's role in inducing aromatase expression, enhanced rates of estrogenesis in developing male frogs, demasculization and feminization of affected individuals, decreased reproductive success, and, finally, population level outcomes, e.g. local extinction or adaptation. The absence of this sort of relatively complete mereological level-hierarchical causal chain would imply "black boxes" that would potentially frustrate attempts to explain higher level phenomena in terms of atrazine exposure.

Third, (similar to the third of Love's criteria for explanations of innovation and novelty), adequate answers to questions about atrazine's endocrine disrupting effects on amphibians in the wild must be constrained by considerations of generalization. There seem to be two dimensions of generalization at play here. The first concerns inferring the presence of mechanisms of endocrine disruption (e.g., aromatase induction) in a given clade or clades from the presence of such mechanisms in another clade or clades. The second concerns generalizing from the (biochemical, cellular, organismal, or populational) effects of endocrine disruption in one clade to similar effects in another. With respect to the first dimension, CYP19 aromatase induction due to atrazine exposure seems to be a mechanism conserved across the vertebrate classes, so here generalizations from one amphibian clade to others seem appropriate. Similarly, aromatase catalyzation of estrogenesis appears

to be highly conserved (Hayes 2005). With respect to the second dimension, can we infer from population level effects of atrazine in one amphibian clade to similar effects in another? In this case, perhaps not, because sex-steroid mediated developmental endpoints may differ among clades (Hayes 2005), and so population level effects will also be likely to differ.

8. Evolutionary biology

I will now use the above proposed criteria to take up a question that was posed at the outset: what role does evolutionary biology play in research on the ecological effects of atrazine as an amphibian endocrine disruptor? First, evolutionary biology can provide population genetic models of amphibian populations, e.g. models of sex ratios in amphibian clades. These hypothetical populations provide null hypotheses (or baseline characterizations) against which claims of atrazine impact can be tested. This contribution of evolutionary biology is disclosed by consideration of the "compared to control" criterion of characterizational adequacy.

Second, evolutionary and (evolutionary developmental) biology provides models of relations among levels of biological organization. Love says that such relations can be understood spatially and temporally both in ontogeny and evolution. Temporal hierarchies in development articulate the relation of, e.g., gene expression to the formation of physiological pathways and morphological structures (2008, 880). In the atrazine case, developmental endocrinology explains how sex steroids at the biochemical level determine the development of sex-specific traits in amphibians at the organismal level. Evolutionary biology contributes here by

providing models linking such traits to population level phenomena; population genetic models can articulate relations between organismal traits and population level effects e.g., covariance between abnormal sex ratios as a result of atrazine-induced feminization (aggregated from the sexual character states of individual organisms) and mean fitness in amphibian clades (Hayes 2010; Guiterres and Teem 2006).

Finally, evolutionary biology contributes phylogenies of relevant traits, e.g. phylogenies of the CYP19 gene, the sex steroids and their receptors, and phylogenies of certain developmental pathways that are mediated by these steroids. Together, these phylogenies are informative about the degree to which atrazine generalizes as an endocrine disruptor and what its likely effects are across diverse amphibian clades. These phylogenies play an important role in satisfying the "generalization" criterion because such phylogenies can either justify or proscribe inferences from research on one clade to claims about another. Importantly for human health, aromatase induction and its effects on sex steroids appear to be conserved across vertebrates (Hayes 2005).

8. Conclusion

Here I've used Love (2008)'s problem agenda framework to characterize research on the impact of atrazine's endocrine disrupting effects on amphibians as addressing a question located within the problem of assessing the impacts of endocrine disruptors in the environment.

This problem is seen as shared by the problem agendas of environmental toxicity and developmental endocrine function. To characterize the epistemic goal

of impact assessment central to the environmental problem of endocrine disruptors, I have developed and deployed the concept of criteria of characterizational adequacy, constraints of adequacy on empirically-grounded characterizations and the processes that generate them. This concept, along with Love (2008)'s concept of criteria of explanatory adequacy, make clearer the ways in which various disciplines make their contributions to the problem of atrazine toxicity and the question of atrazine's endocrine disrupting effects on amphibians. In particular, we've seen that evolutionary biology contributes by providing models of relevant evolutionary processes and phylogenies that inform the propriety of generalizing from findings about one clade to claims about others. Evolutionary biology also contributes by providing models of population-level phenomena that may result from organismal-level atrazine exposure effects.

The forgoing treatment of atrazine research can be seen as a further development of Love (2008)'s and Brigandt (2010)'s response to the challenge issued by Rosenberg (1997) and others. This challenge is for those who participate in the skeptical consensus about the prospects and motivation for Nagelian-type theory reduction to provide alternative accounts of the epistemic relations among scientific disciplines. Love and Brigandt have provided nonreductionist accounts of disciplinary integration centered on solving particular problems and providing particular explanations in evo-devo. Here we've seen how Love's problem agenda framework can be applied to another area of research by expanding this framework to include criteria of characterizational adequacy, criteria constraining what counts as an adequate empirically-grounded characterization given the problems that such

characterizations are meant to address. In this way, the forgoing treatment of atrazine research is meant to provide a modest contribution the broader project of giving plausible nonreductionist problem-centered philosophical accounts of the epistemic relationships among scientific and especially biological disciplines.

Bibliography

Appleby, A. P.; Muller, F.; Carpy, S. (2001), *Weed control in agrochemicals*. Wiley-VCH: New York.

Brigandt, Ingo (2010), "Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology", *Erkenn* 73:295–311

Brigandt, Ingo and Love, Alan (2012) "Reductionism in Biology", *The Stanford Encyclopedia of Philosophy (Summer 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/reduction-biology/>.

Dupré, J. (1993), *The disorder of things: metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Gutierres JB, Teem Jl (2006), "A model describing the effect of sex-reversed YY ␣ sh in
an established wild population: The use of a Trojan Y chromosome to cause extinctionof an introduced exotic species *" J Theor Biol* 241:333–341.

Hayes, Tyrone (2005), "Welcome to the revolution: Integrative Biology and Assessing the Impact of Endocrine Disruptors on Environmental and Public Health. *Integr. Comp. Biol*., 45:321–329

Hayes *et al.* (2010), "Atrazine induces complete feminization and chemical castration in male African clawed frogs(Xenopus laevis)" *PNAS* March 9, 2010 vol. 107 no. 10 4612-4617

Hayes *et a*l. (2011), "Demasculinization and feminization of male gonads by atrazine: consistent effects across vertebrate classes", *J Steroid Biochem Mol Biol* 127:64 –73

Kellert, S.H., H.E. Longino, and C.K. Waters (2006), "Introduction: the pluralist stance", in S.H. Kellert, H.E. Longino, and C.K. Waters (eds.), *Scientific pluralism*

(Minnesota Studies in Philosophy of Science, Vol. 19), Minneapolis: University of Minnesota Press, vii– xxix.

Kitcher, P. (2001), *Science, truth and democracy*. Oxford: Oxford University Press

Landis, W. G., and Yu, M. H. (2010), *Introduction to environmental toxicology: Impacts of chemicals upon ecological systems*. Taylor: Boca Raton, Florida.

Love, Alan (2008), "Explaining evolutionary innovations and novelties: Criteria of explanatory adequacy and epistemological prerequisites," *Philosophy of Science*, 75: 874–886.

Nagel, E. (1961), *The structure of science*. New York: Harcourt, Brace, and World.

Robert, J.S. (2004), *Embryology, epigenesis, and evolution: taking development seriously*. New York: Cambridge University Press.

Rohr, J.R., McCoy K.A. (2010), "A qualitative meta-analysis reveals consistent effects of atrazine on freshwater fish and amphibians", *Environ Health Persp* 118, 20–32

Rohr, J.R., McCoy K.A. (2010), "Preserving environmental health and scientific credibility: a practical guide to reducing conflicts of interest*", Conservation Letters* 3 143–150

Rosenberg, A. (1997), "Reductionism redux: computing the embryo", *Biology and Philosophy* 12:445–470.

Schaffner, K. F. (1969), "The Watson-Crick model and reductionism", *British Journal for the Philosophy of Science,* 20, 325–348.

Schaffner, K. F. (1993), *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.

Tabery, J.G. (2004), "Synthesizing activities and interactions in the concept of a mechanism", *Philosophy of Science* 71:1–15.

Waters, C.K (2007), "The nature and context of exploratory research", *Hist. Phil. Life Sci.* 29(3).

Wimsatt , William (1974), "Reductive explanation: a functional account." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* , Vol. 1974 671-710

Witschi, E. (1929), "Rudimentary hermaphroditism and Y chromosome in *Rana temporaria*", *J. Exp. Zool*. 54:157–223.

22

Imprecise Probability and Higher Order Vagueness

Susanna Rinard
Harvard University
July 10, 2014

Preliminary Draft. Do Not Cite Without Permission.

Abstract

There is a trade-off between specificity and accuracy in existing models of belief. Descriptions of agents in the tripartite model, which recognizes only three doxastic attitudes—belief, disbelief, and suspension of judgment—are typically *accurate*, but not sufficiently *specific*. The orthodox Bayesian model, which requires real-valued credences, is perfectly *specific*, but often *inaccurate*: we often lack precise credences. I argue, first, that a popular attempt to fix the Bayesian model by using sets of functions is also inaccurate, since it requires us to have interval-valued credences with *perfectly precise* endpoints. We can see this problem as analogous to the problem of higher order vagueness. Ultimately, I argue, the only way to avoid these problems is to endorse *Insurmountable Unclassifiability*. This principle has some surprising and radical consequences. For example, it entails that the trade-off between accuracy and specificity is in-principle unavoidable: sometimes it is simply impossible to characterize an agent's doxastic state in a way that is both *fully accurate* and *maximally specific*. What we *can* do, however, is improve on both the tripartite and existing Bayesian models. I construct a new model of belief—the *minimal model*—that allows us to characterize agents with much greater specificity than the tripartite model, and yet which remains, unlike existing Bayesian models, perfectly accurate.

0. Introduction

Much traditional epistemology employs a tripartite model of belief, which recognizes three doxastic attitudes: belief, disbelief and suspension of judgment. However, this model is too coarse-grained. Its descriptions of agents are typically *accurate*, but they are not sufficiently *specific*. For example, one may believe both P1 and P2, and yet be more confident of one than the other. The tripartite model is blind to these differences in confidence, and yet they are crucially important (for example, in the explanation of action).

Observations like these are often taken to motivate the orthodox Bayesian model, which recognizes uncountably many different doxastic attitudes: one for each real number between 0 and 1. However, this model suffers from the opposite problem. It is too fine-grained. Descriptions of agents in this model are perfectly *specific*—reflecting even the subtlest differences in confidence—but this comes at the cost of *accuracy*. Sometimes we lack precise, point-valued credences.

The tripartite model, then, has accuracy without specificity; the orthodox Bayesian model has specificity without accuracy. One aim of this paper is to present powerful reasons for the claim that this trade-off between accuracy and specificity in the

representation of belief is *in principle unavoidable*: it is simply not possible to represent belief in a way that is both *fully accurate* and *maximally specific*.

What we *can* do, however, is improve on both the tripartite and orthodox Bayesian models. Another aim of the paper is to construct and promote a new model of belief—the *minimal model*—which allows for much greater specificity than the tripartite model, and yet which remains, unlike the orthodox Bayesian model, perfectly accurate.

I begin in section 1 by discussing a popular attempt to improve on the orthodox Bayesian model by using a *set* of functions to represent an agent, rather than a *single* function. It will turn out that this set-of-functions model suffers from a slightly different version of the same sort of inaccuracy; specifically, it requires interval-valued credences with *perfectly precise* endpoints. An analogy between credal imprecision and vagueness will reveal that the set-of-functions model is analogous to supervaluationism, and the problem of interval endpoint precision analogous to the problem of higher-order vagueness. We'll see that other approaches to vagueness inspire analogous models of belief, but many suffer from analogs of some version of the higher-order vagueness problem.

In section 2 I argue that the only way to avoid these problems is by endorsing a surprising and radical principle I call Insurmountable Unclassifiability. I then ask what can be said about the representation of belief if this principle is true. As a preliminary to answering, in section 3 I present a novel way of using sets of functions, and intervals, to characterize doxastic states. This *minimal model* allows us to give multiple accurate characterizations of the same agent at different levels of specificity.

The minimal model has two primary virtues. First, it constitutes an improvement over the tripartite model and both existing Bayesian models, in that it allows for much greater specificity than the tripartite model, and yet, unlike existing Bayesian models, remains fully accurate. Second, it is neutral on the issues that divide defenders of different views on credal imprecision and higher order vagueness. Thus, it provides a framework within which we can make progress on a wide range of questions in epistemology and decision theory without having to first take a stand on these difficult and controversial issues.

In section 4 I make use of the minimal model in showing that if Insurmountable Unclassifiability is true, then the trade-off between accuracy and specificity in the representation of belief is *in principle unavoidable*: it is impossible to characterize an agent's doxastic state in a way that is both fully accurate and maximally specific.

Insurmountable Unclassifiability is a radical and mysterious view, however, as becomes apparent in section 5, where I discuss the forced march scenario. In section 6 I display some advantages of the view by showing how it can dissolve two challenging problems: puzzle cases for the Principle of Indifference, and diachronic decision problems for agents who lack precise credences.

Ultimately I do not come down one way or the other on whether we should accept this fascinating view. My aim is only to argue that there is a powerful reason in favor of it—namely, that it is necessary for avoiding the analogous problems of inaccuracy due to overprecision in the representation of belief, and higher order vagueness—and to begin exploring some of its consequences.

1. The Vagueness Analogy

In this section, I begin by spotlighting the problem faced by the orthodox Bayesian model. Then I discuss a popular attempt to fix this problem by using a *set* of functions, rather than a *single* function. I point out that this set-of-functions model faces a problem very similar to the one that undermined the single-function model. Pursuing an analogy between credal imprecision and vagueness helps us gain deeper insight into the nature of these problems. In particular, we can see the set-of-functions model as analogous to the supervaluationist approach to vagueness, and the problem facing this model as analogous to a well-known problem for supervaluationism, sometimes called the *problem of higher-order vagueness*. I then briefly review two other approaches to vagueness: an infinite hierarchy of borderline cases, and many-valued logics. In each case, we see that it would be possible to construct an analogous model of belief, but that the resulting model would suffer from a version of the higher-order vagueness problem.

The orthodox Bayesian model represents the doxastic state of an agent by a function which assigns to each proposition some real number between 0 and 1. However, our confidence levels are not always that precise. For example, consider LUCKY, the proposition that you will find a four-leaf clover tomorrow. I could inquire about your level of confidence in LUCKY, and demand that you choose exactly one real number between 0 and 1, precise down to the millionth decimal place (and beyond)—but there would be an element of arbitrariness in any choice you might make.[1, 2]

One popular attempt to fix this problem involves using a *set* of functions to represent each agent, rather than a *single* function. (Proponents include Jeffrey (1983), van Fraassen (1990), and Joyce (2005, 2010).) This set generates interval-valued credences; one's credence in P is the interval which contains all and only the real numbers r such that some function in one's set has $\Pr(P) = r$.

However, this model faces a problem much like the one that undermined the single-function model. The problem of arbitrariness resurfaces, just in a slightly different form. What, exactly, is the upper endpoint of your interval-valued credence for LUCKY? .0001? .00009? .000121? Again, any particular number seems arbitrary.[3]

We can gain a deeper understanding of these problems of arbitrariness by pursuing an analogy between credal imprecision and vagueness. To begin, I'll define a predicate MC. Roughly, MC applies to numbers that represent credences greater than your level of confidence in LUCKY. For a more precise definition, first, let B[r] say that

[1] Some insist that despite appearances to the contrary, every agent does in fact have a precise credence in every proposition. Arguing against this view is beyond the scope of this paper. Those who are sympathetic to it may interpret the paper as aiming to defend the conditional claim that the conclusions pursued here will follow, *if* in fact we sometimes lack precise credences.

[2] One might object to my presupposition that this inaccuracy in the orthodox Bayesian model is problematic. After all, simplifying idealizations are commonly used with great success in models in science. My reply is simple: whether an idealization is problematic depends entirely on the purpose to which the model is put. For example, if the goal is prediction, an idealization might be unproblematic (and even beneficial). In this paper, however, I evaluate models according to their ability to satisfy a simple curiosity about the actual facts—the truth, the whole truth, and nothing but the truth—about the nature of our doxastic states. Relative to this goal, any inaccuracy in the way a model represents our doxastic states is automatically a shortcoming.

[3] The problem of interval endpoint arbitrariness is also raised (though not under that name) in Sturgeon (2008, 158) and Maher (2006), and discussed in Kaplan (2010).

a coin with bias *r* towards heads will land heads on the next toss. For example, B[.7] says that a coin with bias .7 towards heads will land heads on the next toss. Assume that, in accordance with rationality, for all *r*, your credence in B[r] is *r*. Then we can define MC (for *more confident*) as the predicate that applies to a number *r* just in case you're more confident of B[r] than LUCKY.

MC is analogous to paradigm vague predicates like TALL. In each case there is a spectrum such that the predicate clearly applies at one end, and clearly fails to apply at the other, but there does not seem to be a sharp boundary. Both seem to admit of borderline cases, and give rise to sorites-style paradoxes.

Importantly, there is a close connection between our lack of a precise credence in LUCKY and the vagueness of MC.[4] If we had some precise credence *c* in LUCKY, there would be a sharp boundary between MC and not-MC: all numbers greater than *c* would have MC, and all numbers equal to or less than *c* would have not-MC. A natural thought, then, is that credal imprecision is of a piece with the vagueness of predicates like MC.

The set-of-functions model is an attempt to account for our lack of precise credences, and, thereby, an attempt to account for the vagueness of predicates like MC. We can interpret this model along supervaluationist lines: each function in your set is one admissible precisification of your doxastic state. Functions excluded from your set are *in*admissible precisifications. A proposition about your doxastic state is *determinately* true if true according to all functions in your set. It's *indeterminate* if true according to some, but not all, functions in your set.

For example, if all functions in your set have Pr(B[.9]) > Pr(LUCKY), it's determinate that you're more confident of B[.9] than LUCKY, i.e., it's determinate that MC[.9] is true. If some functions in your set have Pr(B[.0001]) > Pr(LUCKY) while others do not, it's indeterminate whether you're more confident of B[.0001]) than LUCKY, i.e., it's indeterminate whether MC[.0001] is true. So, on the supervaluationist interpretation, we can see the set-of-functions model as an attempt to account for the vagueness of MC by introducing a third category, *indeterminately* MC, just as the supervaluationist tries to account for the vagueness of TALL by introducing a third category, *indeterminately* TALL.[5]

Supervaluationism faces a well-known problem: it requires a sharp boundary where, intuitively, there shouldn't be one, namely, between the *determinate* and *indeterminate*. This is the so-called problem of higher-order vagueness. On the supervaluationist interpretation of the set-of-functions model, the problematic requirement of precise endpoints for interval-valued credences is just an instance of this more general problem. A sharp upper endpoint for one's interval-valued credence in LUCKY would constitute a sharp boundary between the numbers to which MC *determinately* applies and those to which it *indeterminately* applies. But there is no sharp boundary here.[6]

[4] For simplicity I will speak as if MC is vague, but strictly speaking I rely only on the claim that MC is structurally analogous to paradigm vague predicates.

[5] Supervaluationist approaches to vagueness are defended in Fine (1975) and Keefe (2000), among others.

[6] Unfortunately, the question of how to interpret the set-of-functions model has received very little attention in the literature, so although some explicitly endorse the supervaluationist interpretation (including van Fraassen (1990, 2005, 2006) and Hajek (2003)), it is unclear how widespread this interpretation is, and what alternatives might look like. One clearly incompatible interpretation appears in Schoenfield (2012) and Kaplan (2010): if some functions in your set have Pr(A) > Pr(B), while others have Pr(A) < Pr(B), and

As the literature has shown, it is extremely difficult to do justice to higher order vagueness. Below I briefly review a couple of prominent attempts and why they are unsatisfactory. Because of the analogy between credal imprecision and vagueness, in each case there is an analogous model of belief that would face analogous problems.

First, some postulate an infinite hierarchy of borderline cases, borderline borderline cases, etc. For example, indeterminately tall heights are first-order borderline cases, but there are also second-order borderline cases: heights that are borderlines of the distinction between determinately tall and indeterminately tall. For every natural number *n*, there are *n*th-order borderline cases. This strategy could be employed in an attempt to fix the problem of precise endpoints for interval-valued credences: we would account for the lack of a sharp line between determinately MC and indeterminately MC by postulating borderline cases of that distinction, and so on up the hierarchy.

However, there is a compelling objection to this approach. Consider paradigm tall heights that aren't borderline cases at *any* level. Call these *absolutely* tall. If every height must be classified as either absolutely tall, or absolutely not tall, or, for some natural number *n*, a borderline case of *n*th order, then there will be a sharp cut-off between those classified as absolutely tall and those classified some other way. But there should not be a sharp cut-off here.[7] Similarly, if every real number between 0 and 1 must be classified as either absolutely MC, or absolutely not MC, or, for some natural number *n*, a borderline case of *n*th order, then there will be a sharp cut-off between the numbers that are absolutely MC and those that are not. But there is no sharp boundary here.

A different class of views on vagueness involves multiple *degrees* of truth.[8] Some postulate uncountably many: one for each number between 0 and 1, inclusive. This view also faces the problem of higher order vagueness. If we assign some precise degree of truth to every proposition of the form TALL(r), there will be a sharp cut-off between the numbers such that TALL(r) is true to degree 1, and the numbers such that TALL(r) is true to some degree less than 1. But there should not be a sharp cut-off here.

The analogous model of belief, on which each proposition of the form MC(r) is assigned some precise degree of truth, suffers from the same problem: it requires a sharp cut-off between those numbers for which MC(r) is true to degree 1, and those for which MC(r) is true to some degree less than 1. But again, there is no sharp boundary here.

2. Insurmountable Unclassifiability

The orthodox Bayesian model of belief, as we have seen, is inaccurate. The nature of the inaccuracy is that it requires point-valued credences, which, as we saw in the previous section, amounts to the requirement that there be sharp lines where, intuitively, there aren't any—such as between MC and not-MC. (MC(r) says, recall, that

---

still others Pr(A) = Pr(B), then it's false that you're more confident of A than B; false that you're less confident of A than B; and false that you're equally confident of A than B. Your attitude towards A and B falls into a fourth category. (This is analogous to a view in value theory defended, among others, by Ruth Chang (see, for example, Chang (2002)), on which it can be the case that A is neither more valuable than B, nor less valuable, nor equally valuable.) This interpretation requires, implausibly, a sharp boundary between the numbers that have MC, and those that fall into the fourth category.

[7] Similar problems are presented in Sainsbury (1991) and Wright (2009).
[8] See, for example, Smith (2008) and Zadeh (1975).

you are more confident of B[r]—that a coin with bias *r* towards heads will land heads on the next toss—than LUCKY (that you will find a four-leaf clover tomorrow).)

So, we cannot use the two categories *MC* and *not-MC* to give an exhaustive classification of all numbers between 0 and 1. In the previous section we surveyed a number of different attempts to reach a categorization that *is* genuinely exhaustive by *refining* this initial two-way distinction. Specifically, each attempt involved adding more (in some cases, infinitely more) intermediate categories. But all these attempts foundered on the same sort of problem: each ended up requiring sharp lines where, intuitively, there shouldn't be any.

In fact, we can identify an even more substantive commonality among the sources of failure of these different attempts. In each case, the proposed model failed because it ended up requiring a sharp line between those numbers for which MC(r) is *as true as possible*, and those for which it is not. For example, consider the supervaluationist framework. Here, to be determinately true is to be as true as possible, and to be indeterminately true is to fail to be as true as possible. So the sharp line, required by this framework, between the determinate and the indeterminate amounts to a sharp line between what's as true as possible and what's not. Postulating an infinite hierarchy of borderline cases doesn't help; we still have a sharp line between what's as true as possible (in this case, when MC(r) is *absolutely* true, not borderline at *any* level) and what's not (when MC(r) is either a borderline case at *some* level, or absolutely false). Many-valued logics require a sharp line between those numbers for which MC(r) is true to degree 1 (as true as possible), and those true to some degree less than 1 (not as true as possible).

The error that unifies these models, then, is their joint commitment to a sharp line between what's true as possible and what's not. That there is no sharp line here imposes a severe restriction on our ability to classify the numbers between 0 and 1 according to their MC status. This restriction can be stated more precisely, as follows:

Insurmountable Unclassifiability: For any set of categories C with properties (1) and (2) (definitions to follow), it is not the case that every number between 0 and 1 can be classified into some category in C. (1) For some proper subset of C, any number r that falls into some category in that subset is, in virtue of being in that category, such that MC(r) is as true as possible. (2) Any number r that falls into some category in C in the complement of that proper subset is, in virtue of being in that category, such that it's not the case that MC(r) is as true as possible.

Insurmountable Unclassifiability is intended to capture the idea that, not only can we not classify all numbers along the two-way MC/not-MC distinction, but we also cannot classify all numbers according to any set of categories that is intended to *supersede* this two-way distinction, or *refine* it via the addition of (even infinitely many) more intermediate categories.[9] What is wrong the initial two-way distinction is not that there are too few categories into which to sort numbers. Rather, the fundamental problem is the exhaustiveness assumption. The mistake is to think that *every* number can be

[9] Sainsbury (1991) sketches a view that has much in common with Insurmountable Unclassifiability.

assigned to some category that completely captures its status with respect to MC. Insurmountable Unclassifiability denies this.[10]

Now, it is clear that *some* numbers are classifiable in categories that are members of a set with properties (1) and (2). What Insurmountable Unclassifiability denies is just that *all* numbers are so classifiable. That some, but not all, numbers are so classifiable might lead one to think that we can draw a sharp line between those that are, and those that aren't. But, as we will see, this, too, is not possible. That it's not begins to illustrate the truly *insurmountable* nature of the unclassifiability.

Suppose we were able to say, for each number, whether it was so classifiable or not. If so, then, for each classifiable number *r*, we can say whether or not MC(r) is as true as possible. (If we couldn't, then *r* wouldn't be in the classifiable category.) So, there would be a sharp two-way distinction between, on the one hand, the numbers that are classifiable, and such that MC(r) is as true as possible; and, on the other, those that are either unclassifiable, or, classifiable, but not as such that MC(r) is as true as possible. But that would constitute a sharp two-way classification between the numbers for which MC(r) is as true as possible, and those for which it is not! Each number in the first category—classifiable, and such that MC(r) is true as possible—is clearly such that MC(r) is true as possible. Of the numbers in the second category, those that are classifiable, but not such that MC(r) is true as possible, are, obviously, not such that MC(r) is true as possible. And—crucially—those in the second category that are not classifiable are clearly such that it's not the case that MC(r) is true as possible. After all, if MC(r) *were* as true as possible, then that number would fall into the first category. So Insurmountable Unclassifiability entails that, although some numbers are classifiable, and it's not the case that all numbers are classifiable, we cannot categorize each number as either classifiable or not.

As will become even more apparent later in the paper, Insurmountable Unclassifiability is quite a radical and mysterious view. In this section I have argued, though, that accepting it is absolutely necessary if we are determined to avoid problems of the sort that plagued the orthodox Bayesian model and the myriad attempts, discussed in the previous section, to improve upon it. That is, accepting Insurmountable Unclassifiability is absolutely necessary to avoid commitment to the existence of sharp lines where, intuitively, there shouldn't be any.

3. The Minimal Model

I began this paper by noting a trade-off between accuracy and specificity in two popular models of belief. In the section following this one I will show that if Insurmountable Unclassifiability is true, then this trade-off is in principle unavoidable: it

[10] Those with supervaluationist inclinations might think the solution is to posit a third category, consisting of borderline cases of the distinction between what's as true as possible and what's not. But there can be no borderlines of this distinction. Suppose there were. Then there would be a distinction between the determinately true as possible, and the borderline true as possible. Consider an instance of the second category. Since it is determinately in the second category, rather than the first, it determinedly fails to be as true as it possibly could be. (It could be in the first category.) So what we would have is not *borderline* as true as possible, but rather, determinately *not* as true as possible. It is part of the nature of the distinction between what's as true *as possible*, and what's not, that there can be no borderlines of it. (Compare the argument in Broome (1997) for his Collapsing Principle and an argument in Barnes (1982), p. 55.)

is impossible to characterize belief in a way that is both fully accurate and maximally specific. In order to do this, I will first construct a new way of using interval notation and sets of functions to represent belief, which I call the *minimal model*. That is the aim of this section. As we will soon see, it turns out that this minimal model constitutes an improvement over both the tripartite model and existing Bayesian models: it allows for much greater specificity than the tripartite model, and yet remains, unlike existing Bayesian models, perfectly accurate.

First, notice that among *informal* characterizations of belief, some are *more specific* than others. For example, suppose I say you're *more confident of P than not*, and then elaborate that you're *nearly certain of P*. The second description is more specific than the first, but both are perfectly accurate. The new *minimal* interpretation of sets and intervals presented here allows us to give, in a similar fashion, multiple accurate characterizations, at different levels of specificity, of a single agent's doxastic state.

I'll start with intervals. First I'll give a definition that is helpful, but potentially misleading; then I'll do it more carefully. Helpful, but misleading: on the minimal interpretation, [c, d] accurately characterizes an agent's doxastic state toward H just in case *the agent's level of confidence in H is contained within [c, d]*. Note that on this interpretation there will always be multiple accurate intervals at varying levels of specificity. On reason is that, if [c, d] is accurate, then any larger (more inclusive) interval is also accurate. This is because, if one's level of confidence is contained within [c, d], then it is contained within any interval of which [c, d] is a subset. For example, since I am fairly confident that LUCKY is false, my doxastic attitude toward LUCKY is accurately characterized by [0, 1], [0, .8], [0, .5], and many others. On the minimal interpretation, each of these is perfectly accurate; some are more specific than others.

The definition given above is misleading insofar as it presupposes that the agent has a precise, point-valued level of confidence. The fix is: on the minimal interpretation, an interval [c, d] accurately characterizes an agent's doxastic state toward H just in case (1) the agent is *more* confident of H than any proposition in which her credence is *less* than *c*; and (2) the agent is *less* confident of H than any proposition in which her credence is *greater* than *d*. One final clarification: the conditions on the right-hand side must be *as true as possible* for the interval to count as accurate.

One noteworthy feature of this new use of intervals is that it is entirely compatible with the substantive views held by proponents of different views on credal imprecision, including the supervaluationist interpretation, an infinite hierarchy of borderline cases, many-valued logics, Insurmountable Unclassifiability, even the single-function model, etc. For example, if the supervaluationist has [c, d] as the agent's unique interval-valued credence in H, then that interval (as well as any more inclusive interval) automatically counts as accurate on the minimal interpretation. This is because the supervaluationist interpretation of the interval notation is logically stronger than the minimal interpretation of the same notation (hence the name *minimal*). For another example, if one has a point-valued credence *c* in H, then every interval of which *c* is a member will count as accurate on the minimal interpretation. What this makes salient is that *to characterize an agent's doxastic attitude using the new, minimal interval notation is to remain neutral on the issues that divide defenders of all different views on credal (im)precision.* This is because, on the minimal interpretation, to characterize an agent's attitude towards H with some interval is to remain completely neutral about the status of numbers *inside* that

interval. For any such number *r*, it may be that one is *more* confident of H than B[r]; or *less* confident; or *equally* confident; it may be indeterminate whether *r* is one's level of confidence in H; etc. For any possible status *r* might have, we remain completely neutral about whether *r* has that status. To describe one's attitude with an interval is to be committal only about numbers outside that interval.

We can give a new, minimal re-interpretation of sets of functions in the same vein. First pass: a set counts as accurate just in case every proposition about the agent's doxastic state that is true according to *all* functions in that set is true (as possible) of the agent. This is only a first-pass definition, though, because some propositions true according to all functions in the set are *not* true of the agent. For example, it is true according to every function in the set that you have a precise credence in LUCKY. (The functions agree that you have a precise credence; they just disagree about what it is.) But *that you have a precise credence in LUCKY* is precisely what we want to deny! We can get to the root of this problem by noticing that, although the existential claim *there is some real number r such that r is your precise credence in LUCKY* is true according to every function, there is no instance of that existential claim that is true according to every function; that is, there is no *r* such that *r is your precise credence in LUCKY* is true according to every function in the set. So, we can fix the problem by revising the definition as follows: First, let Z be the set of all propositions true according to all functions in the set. Generate Z- by removing from Z any proposition that is, or is equivalent to, some existential claim such that no instance of that existential claim is true according to every function in the set. Now the proper definition: on the minimal interpretation, a set counts as accurate just in case every proposition in Z- is true (as possible) of the agent.

To characterize an agent with a set, on this new, minimal interpretation, is to remain entirely neutral on the status of propositions about which different functions in the set disagree, just as the minimal interval notation remains neutral on the status of numbers inside the interval. There will always be multiple accurate sets, just as there are always multiple accurate intervals. In addition, as before, this interpretation is compatible with different views on credal imprecision. For example, if S is the agent's unique set of functions, interpreted in the supervaluationist way, then S (and any more inclusive set) counts as an accurate description of that agent on the minimal interpretation. If an agent is best represented by a single credence function, any set containing that function is accurate.

This minimal model can easily accommodate traditional Bayesian approaches to a wide range of different issues, such as learning from experience, theory confirmation in science, decision theory, etc. It will typically be the case that if, on the single-function model, some proposition B is true of an agent if that agent's credence function has property Q, then, on the minimal model, B will be true of the agent if there is some set of functions, which accurately characterizes that agent, all of whose members have property Q. For example, on the minimal model we can say that E confirms H if there is some accurate set of functions, all of whose members have $Pr(H|E) > Pr(H)$. The agent satisfies minimal synchronic requirements for rationality if there is some accurate set of functions, all of whose members conform to the axioms of probability. The agent rationally updates on evidence E, received between t1 and t2, just any case any set of

functions S2 that is accurate at t2 can be obtained from some set S1 that was accurate at t1 via conditionalizing each function in S1 on E. And so forth.

What I want to emphasize here, though, is that when it comes to the representation of belief, the minimal model is an improvement over both the tripartite model and existing Bayesian models. It allows for characterizations much more specific than those of the tripartite model. For example, if [.95, .96] accurately characterizes my attitude toward A, and [.97, .98] my attitude toward B, the model represents that I am more confident of B than A, even if I believe both. Yet it remains (unlike existing Bayesian models) perfectly accurate. It does not fall prey to the problem of arbitrariness, since accurate intervals are not taken to be *uniquely* accurate. It is perfectly accurate to characterize my attitude toward LUCKY with [0, .8], even though the endpoints are perfectly precise, because this characterization is not taken to be a unique best one. Other intervals, such as [0, .7] or [0, .6], may be equally accurate.

The only shortcoming of this model is that it typically provides only partial descriptions of the agent's doxastic state. To characterize one's doxastic state with [0, .7], for example, is not maximally informative, since narrower intervals may be equally accurate. Whether this shortcoming can be remedied is addressed in the next section.

4. Insurmountable Unclassifiability Renders Impossible the Combination of Perfect Accuracy and Maximal Specificity in the Representation of Belief

In the previous section I constructed the *minimal model*, which allows us to give multiple accurate descriptions of an agent's doxastic state at different levels of specificity. It is fine to characterize one's doxastic state by giving a few descriptions of this kind, but doing so raises a natural question. What is the *most specific* accurate interval? For example, above I listed the following as accurate characterizations of my attitude toward LUCKY: [0, 1], [0, .8], and [0, .5]. Having done this, it is natural to ask about other intervals. What about [0, .3]? (0, .27)? [0, .2]? In particular, it is natural to wonder which interval is the *most specific* interval that is still accurate. After all, any information encoded in a less specific interval is also contained in a more specific interval—so why not just isolate the maximally specific accurate interval, identify it as such, and forget about the rest?

Interestingly, it turns out that if we are serious about avoiding counterintuitive sharp lines—i.e, if we accept Insurmountable Unclassifiability—then we must regard this very natural thought as deeply flawed. Suppose there were a most specific interval that accurately characterized my attitude towards LUCKY—say, [c, d]. Then any narrower interval would fail to be accurate. But then *d* would constitute a precise boundary where there shouldn't be one. For all numbers greater than *d*, MC(r) would be *as true as possible*; for all numbers equal to or less than *d*, MC(r) would fail to be as true as possible. But, as we have seen, Insurmountable Unclassifiability entails that we can't classify all numbers *r* according to this scheme. The upshot: if we accept Insurmountable Unclassifiability—the only way to avoid problems of arbitrariness and higher-order vagueness—then sometimes *there is no maximally specific, fully accurate characterization of one's doxastic state.*

This is a surprising and radical conclusion. It means that it is impossible to give a *complete* description of an agent's doxastic state; there is *no such thing* as a complete

description. We can give partial descriptions of an agent's doxastic state; some will be more specific than others. But we can never identify a particular description as *maximally specific*, or *complete*.[11]

5. Quietism and the Forced March

In order to draw out some further implications of Insurmountable Unclassifiability, consider a *forced march* scenario. Consider the following finite series of intervals. All intervals in the series share the same lower endpoint: 0. The upper endpoint of the first element of the series is 1. For each interval in the series, its upper endpoint is the result of subtracting some minuscule positive real number ε from the upper endpoint of its predecessor. The last element of the series is the first interval whose upper endpoint is equal to or less than 0.

Imagine going through the elements of this series one by one, and asking someone, for each interval, whether it is accurate concerning their level of confidence in LUCKY. At first, the obvious answer is *yes*. But at some point, the response must be something else. (Otherwise the person will answer *yes* to the last element of the series, which is clearly incorrect.) But what can they say?

It follows from Insurmountable Unclassifiability that in order to avoid error, the speaker is at some point required to perform a speech act that is non-committal in the following sense: the content of the speech act is compatible with its being the case that *yes* would have been a correct answer; and also compatible with its being the case that *yes* would have been an *incorrect* answer. Which speech act may have these features is an empirical question. Silence is a natural candidate—but only if it's understood *not* to implicate that the answer is not *yes*. The person might say "I'm bored of this. Let's go swimming!" or "Was that a Pileated Woodpecker that just flew by?"

This suggestion—that, at some point, one must switch to a non-committal speech act—may strike some as unsatisfying. Those who find it unsatisfying may try to stipulate that the subject will say *yes* if and only if *yes* would be a correct answer. Then speech acts like "Let's go swimming!" won't allow them to wriggle out of the question.

However, a defender of Insurmountable Unclassifiability should deny that such a stipulation can be made. On the assumption that, for each question, there is a fact of the matter about whether not the subject responded with *yes*, then the claim that they will say *yes* if and only if *yes* is a correct answer is equivalent to the claim that, for each interval,

[11] How might this affect other issues, such as updating, decision theory, etc.? In a sense, not at all—these are still handled in the minimal model exactly as described in the previous section. But Insurmountable Unclassifiability introduces a new wrinkle: there is now no *guarantee* that there will be a fact of the matter about whether the relevant conditions obtain (although there *may* always be a fact of the matter). For example, on the minimal model we have that *if* there is some accurate set of functions, all of whose members have $Pr(H|E) > Pr(H)$, then E confirms H. And if there is *no* accurate set of functions, all of whose members have $Pr(H|E) > Pr(H)$, then it's not the case that E confirms H. But since, according to Insurmountable Unclassifiability, it's not the case that every set is classifiable as either accurate or not, the door is now open to the possibility that it will not be settled whether E confirms H. It is important to emphasize, however, that Insurmountable Unclassifiability does not *require* that there sometimes fail to be a fact of the matter here. It is compatible with Insurmountable Unclassifiability that there is always a fact of the matter about whether E confirms H. The same goes for other issues, such as rational updating, minimal constraints on synchronic rationality, etc.

there's a fact of the matter about whether it's accurate or not. And the denial of this is a core commitment of Insurmountable Unclassifiability.

What we are left with, then, is a form of Quietism. For any subject of the forced march, it must be the case that there is some pair of adjacent intervals such that they answer *yes* to the question about one but give a different answer to the question about the other. That answer must be non-committal in the sense described above.

This section has only begun to explore the implications of Insurmountable Unclassifiability; many further questions remain. Already, though, we can see that it leads to some surprising, radical, and mysterious conclusions. Those who are serious about avoiding the problems of arbitrariness, higher-order vagueness, and counterintuitive sharp lines are in for some exciting times!

6. Dissolving Problems with Insurmountable Unclassifiability

My goal in this section is to highlight some of the virtues of Insurmountable Unclassifiability by showing how it dissolves two challenging problems: (1) puzzle cases for the Principle of Indifference; (2) diachronic decision problems for agents who lack precise credences.

The much-discussed *Principle of Indifference* (POI) says that if one has no more reason to believe A than B, and no more reason to believe B than A, then one's credence in A should equal one's credence in B.[12] In the orthodox Bayesian model, where credences are point-valued, this famously seems to lead to contradiction in some cases. For example, suppose a factory produces cubes of equal size.[13] You know only that the length (L) of a side of a cube is 2 feet or less. Plausibly, you have no more reason to believe $0 < L \leq 1$ than $1 < L \leq 2$, and vice versa; so, according to POI, you must have the same point-valued credence in each of these propositions, namely, ½. But now notice that your initial information is equivalent to the information that the area (A) of a side is 4 feet or less. Now it seems you have no more reason to believe any one of these four possibilities than any of the others: $0 < A \leq 1$; $1 < A \leq 2$; $2 < A \leq 3$; $3 < A \leq 4$. So, according to POI, you must have the same point-valued credence in each, namely, ¼. But this contradicts the recommendation of the first application of the principle, since $0 < L \leq 1$ is equivalent to $0 < A \leq 1$.

Some (including Joyce (2005) and Weatherson (2007)) have claimed that moving to the set-of-functions model—with the attendant move to interval-valued credences—solves the problem: that we can respect both verdicts of the POI by giving the same (unique, maximally specific) interval-valued credence to each member of the two-celled partition, and the same (unique, maximally specific) interval-valued credence to each member of the four-valued partition. However, as I have shown elsewhere [reference removed for blind review], this is coherent only if one's interval-valued credence in all six is [0, 1]. And, as I have argued elsewhere, (maximally specific) interval-valued credences with such extreme endpoints are not rationally permissible [reference removed for blind review].

This entire set of problems evaporates if Insurmountable Unclassifiability is adopted—specifically, if we give up the presupposition that we must have a maximally

[12] For some recent discussion, see, among others, White (2010) and Huemer (2009).
[13] Example adapted from Van Fraassen (1989).

specific, fully accurate interval-valued or point-valued credence in each proposition. First, we can use the minimal model to re-state the POI in a way compatible with Insurmountable Unclassifiability, as follows: if one has no more reason to believe A than B, and no more reason to believe B than A, then, if one is rational, then if some interval R is an accurate characterization of one's attitude toward A, then that same interval R is also an accurate characterization of one's attitude toward B, and vice versa. This principle does not lead to contradiction, even if we agree that you have no more reason to believe $0 < L \leq 1$ than $1 < L \leq 2$, *and* no more reason to believe any of the following than any other: $0 < A \leq 1$; $1 < A \leq 2$; $2 < A \leq 3$; $3 < A \leq 4$. For example, each of the following intervals can coherently accurately characterize your attitude toward each of these six propositions: [0, .75]; [.12, .8]; [.011, .673]; etc. With Insurmountable Unclassifiability in place, there is no longer any requirement to find a *maximally specific* accurate interval for each proposition—and it is *this* requirement that we should give up in light of the contradiction, not the innocuous principle that when you have no more reason to believe one thing than another, you shouldn't take different doxastic attitudes towards them.

Giving up this requirement also pulls the rug out from under an argument in Elga (2010) for the claim that rationality requires us to have a precise, point-valued credence in every proposition. Elga draws this conclusion after considering a number of different possible decision rules one might use if one had a (unique, maximally specific) interval-valued credence in some proposition. In each case, he argues that the rule is unacceptable. He infers that it is not rational to have a (unique, maximally specific) interval-valued credence in any proposition; and concludes from this that all rational credences are point-valued.

Once Insurmountable Unclassifiability is on the table, it's clear that the final step in Elga's argument is invalid. Elga does not consider the possibility that rationality might allow one to lack a point-valued credence and *also* lack a unique, maximally specific interval-valued credence.[14] I will argue that the decision theory that accompanies Insurmountable Unclassifiability gives the right results about the case on which, Elga claims, the view that we have a (maximally specific) interval-valued credence founders.

Elga supposes that the agent has a maximally specific credence in H of [.1, .8]. He then supposes the agent is offered some bet A, immediately followed by another bet B. If the agent accepts bet A, she will gain \$15 if H is true but lose \$10 if H is false. If she accepts bet B, she will lose \$10 if H is true, but gain \$15 if H is false. The thing to notice is that if the agent accepts *both* bets, she is guaranteed to gain \$5. Of course, rejecting both guarantees \$0. So it would be irrational to reject both bets. (We assume the agent cares only about money, etc.) But, Elga claims, no decision theory for (maximally specific) interval-valued credences can accommodate this. For example, according to one popular rule that has been frequently endorsed in the literature, rejecting bet A is permissible, and rejecting bet B is permissible as well.

Now, suppose that, in accordance with Insurmountable Unclassifiability, the agent in this case lacks a maximally specific interval-valued credence, but that [.1, .8] is an accurate characterization (in the minimal model) of their doxastic attitude. What decision-theoretic commitments would follow? As we have seen, the minimal model remains completely silent on matters about which different functions in some accurate set

[14] It is understandable that he does not consider this view, as it has heretofore not been a salient position in the debate.

disagree. In this case, since [.1, .8] is an accurate interval, there will be some accurate set which contains, for each value *v* in [.1, .8], some function with Pr(H) = *v*. Some functions in that set have the expected value of rejecting bet A greater than accepting, but others have the expected value of accepting greater than rejecting. There is no agreement among the members of this set about whether rejecting is permissible or not. So the minimal model remains completely silent on that question. The same is true for bet B. This is good news—the minimal model does not have the implausible consequence that rejecting each bet is permissible.

However, we might wonder whether we can also do justice to the intuition that rejecting both bets would be impermissible. In fact, we can! Although the minimal model does not prohibit the individual action of rejecting bet A; and does not prohibit rejecting bet B; it *does* prohibit the compound action of rejecting both bets. This is because every function in that accurate set agrees that the expected value of rejecting both bets is $0, and that the expected value of accepting both bets is $5; and so all functions agree that rejecting both bets is rationally impermissible. The minimal model yields exactly the result that Elga claimed no decision theory for non-precise credences could accommodate.[15]

7. Conclusion

I began by noting a trade-off between accuracy and specificity in two commonly-employed models of belief. The tripartite model, which recognizes only three doxastic attitudes, allows us to characterize belief in a way that is *perfectly accurate*, but *not sufficiently specific*. The model is too coarse-grained. The orthodox Bayesian model, on the other hand, is too fine-grained. It represents agents in a way that is highly specific, but typically inaccurate. It requires point-valued credences in every proposition, but usually our doxastic attitudes are not this precise (nor does rationality require them to be). A popular modification of the Bayesian model—the set-of-functions model—suffers from a similar sort of problem. It requires a unique interval-valued credence, with perfectly precise endpoints, for each proposition. But our doxastic attitudes are usually not this precise either (nor does rationality require them to be).

I have shown that if Insurmountable Unclassifiability is true, then this trade-off between accuracy and specificity is in-principle unavoidable: it is simply not possible to characterize belief in a way that is both fully accurate and maximally specific. I gave a powerful reason in favor of this principle: it is, I argued, non-negotiable for those serious about avoiding the problem of inaccuracy due to overprecision in the representation of belief, and the analogous problem of higher order vagueness. Moreover, it can dissolve two challenging problems: puzzle cases for the Principle of Indifference, and diachronic decision problems for agents who lack point-valued credences.

[15] One thing this discussion reveals is that Elga's argument against some other decision rules is too quick. For example, what he calls permissive rules have the same verdict as the minimal model—that the compound action of rejecting both bets is impermissible—for similar reasons. However, this is complicated by the fact that according to these rules, rejecting bet A is permissible, as is rejecting bet B. So they are committed to the implausible result that a certain compound action is impermissible even though each component action is permissible. The combination of Insurmountable Unclassifiability and the minimal model has no such implausible commitment.

However, it has not been my aim in this paper to reach a final verdict on Insurmountable Unclassifiability. It will doubtless be controversial; some of its consequences are surprising and radical. For example, it entails that the subject in a forced march scenario is required, at some point, to respond in a way that is maximally non-committal on the question asked.

Regardless of whether or not we accept this fascinating and mysterious view, we can improve on both the tripartite model and the existing Bayesian models of belief. One aim of the paper has been to describe and promote a new model of belief—the *minimal model*—that enables us to use intervals, and sets of functions, to characterize belief in a way that is much more specific than the tripartite model, and yet which does not fall prey to problems of overprecision and so is perfectly accurate. A further virtue of the minimal model is that it remains neutral on the issues that divide defenders of different approaches to the problem of higher order vagueness. Thus, it provides a framework within which we can make progress on a wide range of different issues in epistemology and decision theory without first having to take a stand on this challenging and controversial problem.

References:

Barnes, J. (1982) Medicine, Experience, and Logic. In J. Barnes, J. Brunschwig, M. F. Burnyeat, and M. Schofield (eds.) Science and Speculation. Cambridge: Cambridge University Press.

Broome, J. (1997) Is Incommensurability Vagueness? In R. Chang (ed.) Incommensurability, Incomparability, and Practical Reason. Cambridge: Harvard University Press.

Chang, R. (2002) The Possibility of Parity. Ethics 112: 659-88.

Fine, K. (1975) Vagueness, Truth and Logic. Synthese 30: 265-300.

Hajek, A. (2003) What Conditional Probability Could Not Be. Synthese 137(3): 273-323.

Huemer, M. (2009) Explanationist Aid for the Theory of Inductive Logic. British Journal for the Philosophy of Science 60(2): 345-375.

Jeffrey, R. C. (1983) Bayesianism with a Human Face. In J. Earman (ed.) Minnesota Studies in the Philosophy of Science, volume 10, Testing Scientific Theories. Minneapolis: University of Minnesota Press.

Joyce, J. (2005) How Probabilities Reflect Evidence. Philosophical Perspectives, 19(1): 153-178.

Joyce, J. (2010) A Defence of Imprecise Credences in Inference and Decision Making. Philosophical Perspectives, 24(1): 281-323.

Kaplan, M. (2010) In Defense of Modest Probabilism. Synthese, 176(1):41-55.

Keefe, R. (2000) Theories of Vagueness. Cambridge: Cambridge UP.

Maher, P. (2006) Book Review: David Christensen. *Putting Logic in its Place: Formal Constraints on Rational Belief*. Notre Dame Journal of Formal Logic 47(1): 133-149 (2006).

Sainsbury, M. (1991) Is There Higher-Order Vagueness? Philosophical Quarterly 41(163):167-182.

Schoenfield, M. (2013) Permission to Believe: Why Permissivism is True and What it Tells Us about Irrational Influences on Belief. Nous 47(1).

Schoenfield, M. (2012) Chilling Out on Epistemic Rationality. Philosophical Studies 158(2): 197-219.
Smith, N.J.J., 2008. Vagueness and Degrees of Truth. Oxford: Oxford University Press.
Sturgeon, S. (2008) Reason and the Grain of Belief. Noûs 42(1):139-165.
van Fraassen, B. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.
van Fraassen, B. (1990) Figures in a Probability Landscape. In M. Dunn and A. Gupta (eds.), Truth or Consequences. Dordrecht: Kluwer.
van Fraassen, B. (2005) Conditionalizing on Violated Bell's Inequalities. Analysis, 65(1):27-32.
van Fraassen, B. (2006) Vague Expectation Value Loss. Philosophical Studies 127(3).
Weatherson, B. (2007) The Bayesian and the Dogmatist. Proceedings of the Aristotelian Society, 107: 169 – 85.
White, R. (2014) Evidence Cannot Be Permissive. In M. Steup, J. Turri, and E. Sosa (eds.) Contemporary Debates in Epistemology, Second Edition p. 312-324.
White, R. (2010) Evidential Symmetry and Mushy Credence. In T.S. Gendler and J. Hawthorne (eds.) Oxford Studies in Epistemology, volume 3. Oxford, UK: Oxford University Press.
Wright, C. (2009) The Illusion of Higher-Order Vagueness. In R. Dietz and S. Moruzzi (eds.), Cuts and Clouds. Vaguenesss, its Nature and its Logic. Oxford University Press.
Zadeh, L. (1975) Fuzzy Logic and Approximate Reasoning. Synthese, 30: 407-428.

Curie's hazard: From electromagnetism to symmetry violation

Bryan W. Roberts
August 21, 2014

Abstract. We explore the facts and fiction regarding Curie's own example of Curie's principle. Curie's claim is vindicated in his suggested example of the electrostatics of central fields, but fails in many others. Nevertheless, the failure of Curie's claim is still of special empirical interest, in that it can be seen to underpin the experimental discovery of parity violation and of CP violation in the 20th century.

1. Introduction

Curie (1894) wrote that, "when certain causes produce certain effects, the elements of symmetry of the causes must be found in the produced effects" (Curie 1894, pg. 394)[1]. This claim has received mixed reviews. Brading and Castellani (2013) have suggested that a common interpretation of the principle is faulty, and Norton (2014) has argued that it is an exercise in dubious causal metaphysics. Many have suggested that the principle fails for the phenomenon of spontaneous symmetry breaking in quantum field systems, although Castellani (2003) and Earman (2004) have each argued that this is not the case.

In this paper I'd like to do two things. First, I would like to discuss Curie's own example of his principle in electromagnetism. It is a deceptively simple example. My aim will be to draw out the particular physical facts that allow Curie's statement to succeed in this example, by formulating and proving a sense in which it succeeds, while isolating a sense in which it can also fail. This is the core of what I would like to say: the truth of Curie's statement is contingent on special physical facts, which obtain in some cases but not others.

Second, I would like to point out that one of the more useful applications of Curie's principle is the detection of its failure, which can provide evidence that the laws of nature are symmetry-violating. Many commentators have focused on the connection between Curie's principle and a different concept, that of spontaneous symmetry breaking[2]. Here I will instead point out how Curie's principle played crucial role in

[1] "lorsque certaines causes produisent certains effets, les éléments de symétrie des causes doivent se retrouver dans les effets produits"

[2] C.f. Castellani (2003) and Earman (2004).

the famous experimental detections of parity violation and CP-violation in the 20th century.

2. Curie's example

Curie's original statement is slightly different from the statement that philosophers and physicists have come to refer to as Curie's principle. I will discuss the latter in more detail in the next section. For now, to keep track of the difference, I will refer to Curie's original statement as:

> *Curie's Hazard.* A symmetry of the causes must be a symmetry of the effects.

Curie gave the following example of this hazardous conjecture. Consider two oppositely charged plates, placed close together and centered on an axis as in Figure 1. Think of the charges as a "cause" whose "effect" is to give rise to an electric field. That effect, Curie says, must exhibit all the symmetries of the cause. So, since the charges are rotationally symmetric, the electric field must be too.

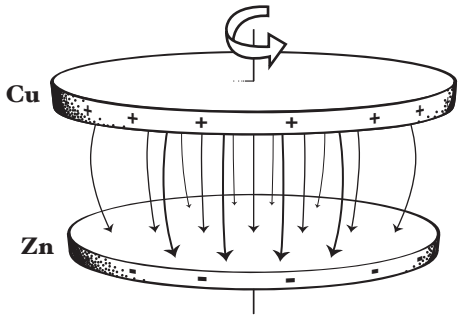


Figure 1. Curie's example: A symmetry of the charges is a symmetry of the electric field.

In Curie's own words:

> To establish the symmetry of the electric field, suppose that this field is produced by two circular plates of zinc and of copper placed one facing the other, like the plates of an air condenser. Considering a point on the common axis between the two plates, we see that this axis is an axis of isotropy and that every plane containing this axis is a plane of symmetry. *The elements of symmetry of the causes must be found in the produced effects; therefore the electric field is compatible with the symmetry* (Curie 1894, pg. 404, emphasis added)[3].

[3]My translation. The original reads: "Pour établir la symétrie du champ électrique, supposons que ce champ soit produit par deux plateaux circulaires de zinc et de cuivre placés en face l'un de

Curie's example is deceptively simple. In the case of two electric plates, it is true that a symmetry of the charge distribution is also a symmetry of the electric field. However, it is not true of Maxwell's equations more generally: a number of implicit assumptions are required in order for it to get off the ground. For example, if the particles in the plates were in motion it would of course cause the field lines to propagate asymmetrically, breaking the symmetry in the charge distribution.

Thus, an obvious implicit premise of Curie's argument must be an absence of motion. But even that is not enough. Suppose that there are no charges at all — that is, consider the vacuum. Maxwell's equations by themselves do not guarantee that an electric field will share the symmetries of the vacuum. On the contrary, there are plane wave solutions to the vacuum Maxwell equations in which the electric field propagates in any direction that one likes (Figure 2).

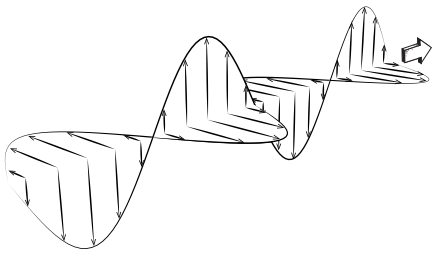


FIGURE 2. The vacuum has no charge or current, but is compatible with an electromagnetic plane wave propagating in any direction.

Getting Curie's example to work thus takes a little bit of care. In the next section, I'll explain how this can be done. If taken truly literally, Curie's hazard is simply wrong: the symmetries of a charge distribution are not necessarily symmetries of the electric field. However, if one presumes a certain amount of special physical facts about particular electromagnetic fields, then there are versions of Curie's hazard that are actually true.

---

l'autre, comme les armatures d'un condensateur à air. Considérons entre les deux plateaux un point de l'axe commun, nous voyons que cet axe est un axe d'isotropie et que tout plan passant par cet axe est un plan de symétrie. Les éléments de symétrie des cause doivent se retrouver dans les effets produits; donc le champ électrique est compatible avec la symétrie" (Curie 1894, pg. 404).

3. A symmetry theorem

Curie's hazard on electromagnetism can be made true given appropriate background assumptions. Let me begin with an informal discussion of the physics underlying Curie's argument, before turning to the more precise formulation of a symmetry principle for electromagnetism along these lines.

3.1. **Physics of Curie's example.**

Curie's two-plate example can be characterized by the following facts.

(1) *Gauss' Law.* Electric fields are related to charge distributions by $\nabla \cdot \mathbf{E} = \rho$.
(2) *Electrostatics.* When there is no change in magnetic field, the electric field is roughly curl-free: $\nabla \times \mathbf{E} = 0$.
(3) *Central field.* The electric field "goes to zero" sufficiently quickly outside of some region, in a sense to be made precise in the next subsection.

These three statements express a divergence and a curl for a vector field that is subject to some appropriate boundary conditions. Given all this, it turns out that Curie's hazard about electric fields holds too. This result stems from two facts; I discuss their proof in the next subsection.

First, it turns out that all of the above relations are preserved by rotations. In particular we have (using $\mathbf{E}'$ and ρ' to represent the rotated field and charges),

(a) *Rotations preserve Gauss' law.* $\nabla \cdot \mathbf{E}' = \rho'$
(b) *Rotations preserve electrostatics.* $\nabla \times \mathbf{E}' = \mathbf{0}$
(c) *Rotations preserve centrality.* $\mathbf{E}' = \mathbf{0}$ on the boundary of and everywhere outside some region.

This kind of reservation does not hold of arbitrary smooth transformation, but does of rotations. We will see shortly that this stems from the fact that a rotation preserves the metric.

Second, an elementary result of vector analysis[4] shows that every central field $\mathbf{v}$ is uniquely determined by its divergence ($\nabla \cdot \mathbf{v}$) and its curl ($\nabla \times \mathbf{v}$). That is, if two such vector fields $\mathbf{v}$ and $\mathbf{v}'$ subject to these boundary conditions have the same divergence and curl, then $\mathbf{v} = \mathbf{v}'$.

These two facts allow one to say why Curie's hazard works in the example of the two plates. Suppose we have a charge distribution that is invariant under rotations:

$$\rho' = \rho.$$

[4]This result is a corollary of what is often called the Helmholtz-Hodge decomposition theorem.

Then by our first observation, the electric field E and its rotated counterpart E' have the same divergence and curl:

$$\nabla \cdot E = \rho = \rho' = \nabla \cdot E'$$
$$\nabla \times E = 0 = \nabla \times E'.$$

But the divergence and curl uniquely determine a vector field under these conditions, so $E' = E$. In other words, when the conditions (1)-(3) are satisfied, then a charge distribution ρ is invariant under rotations only if the electric field E is too, and Curie's hazard is correct.

3.2. **As a general theorem.** The argument above can be stated in more general and rigorous terms as follows. Let M be a smooth manifold, and let g_{ab} be a metric, assumed here to be a smooth symmetric invertible tensor field with inverse g^{ab}; I use Penrose abstract index notation for raising and lowering indices. Let ∇ be the derivative operator compatible with g_{ab} in the sense that $\nabla_a g_{bc} = \mathbf{0}$. A diffeomorphism $\varphi : M \to$ is called an *isometry* if $\varphi_* g_{ab} = g_{ab}$, and is the natural notion of a symmetry in this context.

We begin by collecting two facts about the derivative operator ∇; the proofs are included in an appendix. The first is that isometries "preserve" the derivative operator:

Proposition 1. *If $\varphi : M \to M$ is an isometry and λ^{bc}_d an arbitrary tensor field, then $\varphi_*(\nabla_a \lambda^{bc}_d) = \nabla_a \varphi_* \lambda^{bc}_d$.*

The second fact expresses the sense in which the divergence and the curl uniquely determine a vector field. I will state a geometric version of the standard result, which applies in many more geometries beyond the standard Euclidean metric on $\mathbb{R}^3$. To state this fact, we'll first need a general formulation of the divergence and curl of a vector ξ^a on a 3-dimensional manifold (a 3-manifold) M with volume element[5] ϵ^{abc}:

$$\mathrm{div}(\xi) = \nabla_a \xi^a$$
$$\mathrm{curl}(\xi) = (\nabla \times \xi)^c = \epsilon^{abc} \nabla_a \xi_b.$$

A few more definitions are needed. A metric g_{ab} on M is called *positive definite* if $\xi^a \xi_a \geq 0$, in which case (M, g_{ab}) is called a *Riemannian manifold*. Finally, we define what we mean for a vector field to be a "Central Field":

(Central Field) $E^a = \mathbf{0}$ on the boundary and outside of a region R.

[5] A *volume element* for an n-dimensional Riemannian manifold (M, g_{ab}) is a smooth n-form that satisfies $\epsilon^{a_1 \ldots a_n} \epsilon_{a_1 \ldots a_n} = n!$. A manifold is *oriented* if it admits a volume element.

This formulation is slightly stronger than is necessary. In particular, central fields can be formulated for regions without boundary, such as $\mathbb{R}^3$, a version of our next Proposition still holds so long as the fields go to zero quickly enough (see e.g. Arfken 1985, §1.15). But this minor generalisation is considerably more complicated to state and prove, and the following result is sufficient for our purposes.

Proposition 2. *Let (M, g_{ab}) be an oriented simply-connected 3-dimensional Riemannian manifold, and let ξ^a and χ^b be two vector fields that each satisfy the Central Field assumption with respect to some (possibly different) region. If* $\mathrm{div}(\xi) = \mathrm{div}(\chi)$ *and* $\mathrm{curl}(\xi) = \mathrm{curl}(\chi)$*, then $\xi^a = \chi^a$.*

With these two facts in place, we can now finally state a theorem that captures some general conditions under which Curie's hazardous conjecture is true. The slightly stronger-than-necessary "Central Field" assumption will be adopted here too, as it is simpler and sufficient for our needs.

Theorem. *Let ρ be a scalar field, E^a a vector field, and let $\varphi : M \to M$ be an isometry on an oriented simply-connected 3-dimensional Riemannian manifold (M, g_{ab}). If,*

(1) *(Gauss' law) $\rho = \nabla_a E^a$*
(2) *(Electrostatics) $(\nabla \times E)^a = \mathbf{0}$*
(3) *(Central Field) $E^a = \mathbf{0}$ on the boundary and outside of a region R*

then $\varphi_* \rho = \rho$ (symmetric cause) only if $\varphi_* E^a = E^a$ (symmetric effect).

Proof. Let $\varphi_* \rho = \rho$. By Gauss' law,

$$\nabla_a E^a = \rho = \varphi_* \rho = \varphi_* (\nabla_a E^a) = \nabla_a \varphi_* E^a,$$

where the last equalty is an application of Proposition 1. Thus, E^a and $\varphi_* E^a$ have the same divergence. Moreover, by the assumption of electrostatics,

$$(\nabla \times E)^a = \mathbf{0} = \varphi_* \mathbf{0} = \varphi_* (\nabla \times E)^a.$$

Applying the definition of the curl to the right hand side now gives us,

$$\begin{aligned}(\nabla \times E)^a &= \varphi_* \epsilon^{bca} \nabla_b E_c = \pm \epsilon^{bca} \varphi_* \nabla_b E_c = \pm \epsilon^{bca} \nabla_b \varphi_* E_c \\ &= \pm (\nabla \times \varphi_* E)^a.\end{aligned}$$

where the second equality applies the fact that isometries preserve volume elements up to a sign[6]. But $(\nabla \times E)^a = \mathbf{0}$, so this implies that E^a and $\varphi_* E^a$ have the same curl. Moreover, since E^a is a Central Field with respect to the regions R_i, so is $\varphi_* E^a$ with

[6]Since isometries preserves the metric, $(\varphi_* \epsilon^{bca})(\varphi_* \epsilon_{bca}) = \epsilon^{bca} \epsilon_{bca} = \pm n!$. Thus, $\varphi_* \epsilon^{bca}$ is a volume element too. But ϵ^{bca} and $-\epsilon^{bca}$ are the unique volume elements, so $\varphi_* \epsilon^{bca} = \pm \epsilon^{bca}$.

respect to the regions $\varphi(R_i)$. Therefore, the premises of Proposition 2 are satisfied, and it follows that $E^a = \varphi_* E^a$. □

What I would like to emphasize about this result is that *even for Curie's own example*, the truth of Curie's hazard depends on a significant amount of background structure. It is not an a priori fact about causes and effects. Indeed, the argument does not go through in the more general context of pseudo-Riemannian manifolds, as are adopted in general relativity, except when restricted to a spacelike hypersurface where the metric g_{ab} becomes positive definite. In particular, the proof of Proposition 2 makes crucial use of the non-degenerate metric available in Riemannian manifolds. I do not know if this theorem can be generalized to the pseudo-Riemannian case, but if it can, then it would likely be established by a rather different argument.

3.3. **General electromagnetic fields.** Curie's hazard fares worse when applied to general electromagnetic fields in spacetime. The natural analogue of Curie's statement for electrostatics simply fails when translated into this language.

Electromagnetism is naturally formulated on a smooth manifold M with with a metric g_{ab} that is symmetric and invertible, but not necessarily non-degenerate. Such a pair (M, g_{ab}) is called a *pseudo-Riemannian manifold*. To do electromagnetism, we assume the existence of a vector field J^a representing charge-current density, and an anti-symmetric tensor field F_{ab} representing the electromagnetic field, which satisfy Maxwell's equations[7]:

$$\begin{aligned} \nabla_{[a} F_{bc]} &= \mathbf{0} \\ \nabla_a F^{ab} &= J^b. \end{aligned} \tag{1}$$

These general equations reduce in certain contexts to the usual Maxwell equations (for an overview see Malament 2012, §2.6).

What is Curie's hazard in this context? If we take the cause to be the charge-current J^a (instead of just the charge ρ) and take the effect to be the electromagnetic field F_{ab} (instead of just the electric field E^a), then Curie's statement would be that a symmetry of the charge-current J^a is a symmetry of the electromagnetic field F_{ab}. That statement is false.

The problem is that the charge-current J^a does not uniquely determine an electromagnetic field F_{ab} up to isometry without further specification. This makes it possible to find explicit counter-examples to Curie's hazard, such as the following.

[7] In fact, an even more general formulation is available in terms of the Hodge star operator (for an overview see Baez and Muniain 1994, §1.5), although this will not concern us here.

Counterexample. Let F_{ab}, J^a be a solution to Maxwell's equations, and let $\varphi : M \to M$ be an isometry that does not preserve F_{ab},

$$\varphi_* F_{ab} - F_{ab} = H_{ab} \neq \mathbf{0},$$

but such that H_{ab} is divergence-free, $\nabla_a H^{ab} = \mathbf{0}$. For example, this occurs when F_{ab} is the field for a plane wave (as in Figure 2) and φ is a rotation; then $\nabla_a F^{ab} = \mathbf{0}$ and so $\nabla_a H^{ab} = \mathbf{0}$, but $H_{ab} \neq \mathbf{0}$. Since diffeomorphisms preserve the zero vector, this implies,

$$\varphi_* J^b = \mathbf{0} = J^b.$$

Thus, if J^b is the "cause" and F_{ab} the "effect," then a symmetry of the cause fails to be a symmetry of the effect. Without specifying some initial and boundary conditions such as those considered in the previous subsection, a symmetry of J^b need not be a symmetry of F_{ab}.

A persistent believer in Curie might still draw a more optimistic conclusion. It is easy to see that the converse expression of Curie's claim is true. Suppose we consider F_{ab} to be the "cause" and J^b the "effect", and let $\varphi_* F_{ab} = F_{ab}$. Then applying Proposition 1 we have,

$$\varphi_* J^b = \varphi_*(\nabla_a F^{ab}) = \nabla_a \varphi_* F^{ab} = \nabla_a F^{ab} s = J^b.$$

So, every symmetry of the electromagnetic field F_{ab} is a symmetry of the charge-current field J^a. One could conclude from this that Curie simply mistook cause for effect: the appropriate cause in this example is the electromagnetic field F_{ab}, and the appropriate effect the charge-density current J^a. I do not know what would justify this kind of conclusion; on the contrary, the argument of Norton (2014) suggests it would be little more than dubious causal metaphysics.

Another possible route is to adopt initial and boundary conditions that guarantee J^a will determine a unique electromagnetic field F_{ab}. This is not so easy to do. Wald (1984, Chapter 10, Problem 2) points out some (fairly restrictive) circumstances under which F_{ab} is unique, by demanding that $J^a = \mathbf{0}$, and also that the values of the electric and magnetic fields be on a Cauchy surface. Under these circumstances, one has for *any* isometry $\varphi : M \to M$ that $\varphi_* J^a = J^a = \mathbf{0}$, and so,

$$\nabla_a \varphi_* F^{ab} = \varphi_* \nabla_a F^{ab} = \varphi_* J^b = \mathbf{0} = J^b = \nabla_a F^{ab}.$$

Thus, since F_{ab} is the unique field satisfying Maxwell's equations under these circumstances, it follows that $\varphi_* F_{ab} = F_{ab}$ for all isometries. In other words, Curie's hazard is made true, in that a symmetry of J^a is a symmetry of F_{ab}, in the restrictive circumstances of $J^a = \mathbf{0}$ when there is a complete absence of charge-current in spacetime.

But this is only possible in the presence of this or some similar initial and boundary conditions that render F_{ab} unique.

The point I would like to make about all this is not that Curie's hazard is totally misguided, but rather that very specific structures must be in place for it to be true. Without a number of particular background facts, Curie's hazard can fail, even in his own example of electromagnetism.

4. From electromagnetism to symmetry violation

In this section, I will identify a sense in which the failure of Curie's hazard can provide evidence of symmetry violation. In fact, its failure provides an indicator of symmetry violation in some of the most famous historical examples of symmetry violation: the parity violation detected by Chien-Shung Wu in 1956, and the CP-violation detected by Val Cronin and James Fitch in 1964. I will also now turn to the statement that philosophers and physicists have more commonly come to call "Curie's principle."

4.1. Curie's principle in skeletal form.

The statement known as "Curie's principle" can be cast in an incredibly general form. Here is how to get there from the example of electromagnetism. Under very particular circumstances, a symmetry of the charge-current distribution is also a symmetry of the electromagnetic field. For Curie's two-plate capacitor, we have seen that sufficient circumstances are the electrostatics of central fields: (1) Gauss' law, (2) Electrostatics, and (3) Central Field. Let me now summarise these properties as the statement that the relation between cause and effect is "symmetry preserving," formulated as follows.

Proposition 3 (Curie Principle)**.** *Let C and E be two sets, and let $\sigma_c : C \to C$ and $\sigma_e : E \to E$ be two bijections. If $D : C \to E$ is a mapping such that,*

(symmetry preservation) $D\sigma_c^{-1}x = \sigma_e^{-1}Dx$ for all $x \in C$,

then $\sigma_c x = x$ (symmetric cause) only if $\sigma_e Dx = Dx$ (symmetric effect). If D is a bijection, then $\sigma_c x = x$ if and only if $\sigma_e Dx = Dx$.

Proof. If $\sigma_c x = x$, then $\sigma_e Dx = \sigma_e D(\sigma_c^{-1}x) = \sigma_e(\sigma_e^{-1}D)x = Dx$. If D is a bijection then it has an inverse, so $\sigma_e Dx = Dx$ only if $x = (D^{-1}\sigma_e^{-1}D)x = D^{-1}(D\sigma_c^{-1})x = \sigma_c^{-1}x$ and hence $\sigma_c x = x$. □

The 'σ's are to be interpreted as "the same" symmetry[8], such as a fixed rotation (or whatever), applied to each of the sets C and E. The 'D' mapping captures a sense

[8]One may wish to cash this out as Norton (2014) does in terms of an isomorphism that carries σ_c to σ_e. Or (as is now fashionable) one may interpret this as meaning that C and E are two categories related by a functor $\mathfrak{F}$ with σ_c a morphism of C and σ_e a morphism of E satisfying $\mathfrak{F}(\sigma_c) = \sigma_e$. The

in which "causes determine effects." Note that this formulation explicitly excludes "time reversing" symmetries like T and CPT, since they are typically expressed as mappings between causes and effects[9].

Proposition 3 is similar to some existing formulations of Curie's principle[10]. One sense in which it differs slightly is that it does not presume causes *uniquely* determine effects. When they do not, then the converse statement need not be true, that a symmetry of an effect is necessarily a symmetry of the cause. Elena Castellani[11] has emphasized out that Curie himself viewed his principle as asymmetric:

> In practice, the converse... [of Curie's hazards] are not true, i.e., the effects can be more symmetric than their causes. Certain causes of asymmetry might have no effect on certain phenomena (Curie 1894)[12]

Proposition 3 captures this asymmetry: if causes do not bijectively determine effects, then the converse of Curie's hazard is not guaranteed. However, if a cause does (bijectively) determine a unique effect that satisfies symmetry preservation, then Curie's hazard is true in both directions.

The bare-bones construal of Curie's principle of Proposition 3 can be applied in all kinds of ways. Here are a few, limited only by the imagination.

Example 1 (Electromagnetism)**.** We have already seen the example in which C be is the set charge distributions, and E the set of electric fields on a Riemannian manifold. Here is another way to look at it (which is essentially just the proof of the theorem). Given conditions (1)-(3), Proposition 2 provides a mapping $D : \rho \mapsto E^a$ that determines a unique E^a for each ρ. Proposition 1 then implies[13] that $D\varphi^* = \varphi^* D$ for any isometry φ. Therefore, given (1)-(3), a symmetry of the charge distribution $\varphi_* \rho = \rho$ is also a symmetry of the electric field $\varphi_* E^a = E^a$.

Example 2 (General Relativity)**.** Let C and E both refer to the set of symmetric 2-place tensor fields on a relativistic spacetime (M, g_{ab}), thinking of an element $T_{ab} \in C$ as energy-momentum and an element $G_{ab} \in E$ is the Einstein tensor. Let $D : T_{ab} \mapsto$

difficulty is that "spurious" identifications of symmetries may still occur, as identified by Norton (2014, fn.4).

[9]This is required in order to get a true principle; for time-reversing symmetries, Curie's principle badly fails (Roberts 2013a).

[10]C.f. Ismael (1997), Belot (2003), Earman (2004, pg.175-176), Mittelstaedt and Weingartner (2005, pg.231), Ashtekar (2014) and Norton (2014).

[11]Personal communication.

[12]Translation from Brading and Castellani (2003, pg.312).

[13]Namely, Proposition 1 implies that if $\rho = \nabla_a E^a$ (and hence that $D\rho = E^a$), then $\varphi^* \rho = \varphi^* \nabla_a E^a = \nabla_a \varphi^* E^a$, and thus $D\varphi^* \rho = \varphi^* E^a = \varphi^* D\rho$.

$G_{ab} = \frac{8\pi G}{c^4} T_{ab}$ be the map determined by Einstein's equation. The symmetries of C and E are determined by an underlying diffeomorphism $\varphi : M \to M$, and identifying $\sigma_c = \sigma_e = \varphi_*$ we trivially find that $D\sigma_c = \sigma_e^{-1} D$.

Example 3 (Particle Physics)**.** Let $C = \mathcal{H}_{in}$ and $E = H_{out}$ be identical copies of a Hilbert space $\mathcal{H}$ representing the in-states and out-states of a scattering experiment. Let $D = S : \psi_{in} \mapsto \psi_{out}$ be the scattering matrix. Then the condition that a symmetry σ (a unitary operator) satisfy $D\sigma_{in} = \sigma_{out}^{-1} D$ just amounts to the condition that it commute with the scattering matrix. When this condition obtains, there is a sense in which Curie's hazard is true, that a symmetry of causes (viewed as an "in" state) gives rise to a symmetry of an effect (viewed as an "out" state).

It is this last example that is significant for the history of symmetry violation, to which we will now turn.

4.2. **Curie's failure implies symmetry violation.** When S is a scattering matrix, Proposition 3 says that Curie's hazard holds for a symmetry σ if and only if the S-matrix is "invariant" under that symmetry. For a long time it was presumed that the laws of nature *must* be invariant under symmetries like parity (P) and the combination of charge conjugation and parity (CP). However, in the mid-20th century this presumption was dramatically disproven, when first parity invariance and then CP invariance were both found to be violated in weak interactions.

The significance of Curie's principle for those discoveries can be seen by casting Proposition 3 in the equivalent "contrapositive" form: if Curie's hazard fails, in that either $\sigma_c(x) = x$ and $\sigma_e(Dx) \neq Dx$ or else $\sigma_e(Dx) = Dx$ and $\sigma_c(x) \neq x$, then we have a case of symmetry violation: $D\sigma_c^{-1} \neq \sigma_e^{-1} D$. Applying this principle to particle decays is a little subtle, but not much. That application is stated and proved in Roberts (2013b, Fact 2) as follows.

Proposition 4 (Scattering Curie)**.** *Let S be a scattering matrix, and $R : \mathcal{H} \to \mathcal{H}$ be a unitary bijection. If there exists a decay channel $\psi^{in} \to \psi^{out}$, i.e. a non-zero amplitude $\langle \psi_{out}, S\psi_{in} \rangle$, such that either,*

(1) *(in but not out) $R\psi^{in} = \psi^{in}$ but $R\psi^{out} = -\psi^{out}$, or*
(2) *(out but not in) $R\psi^{out} = \psi^{out}$ but $R\psi^{in} = -\psi^{in}$,*

then,

(3) *$RS \neq SR$.*

This principle is precisely what was used in the very first revelations that the laws of nature are symmetry violating. For example, parity — the "mirror" transformation that reverses total orientation (or "handedness") of a system — has been long known

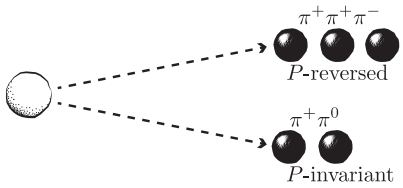


FIGURE 3. The P-violating interaction suggested by Lee and Yang.

to preserve the two-pion state $\pi^+\pi^0$, but reverse the phase of the three-pion state $\pi^+\pi^+\pi^-$:

$$P\pi^+\pi^0 = \pi^+\pi^0$$
$$P\pi^+\pi^+\pi^- = -\pi^+\pi^+\pi^-.$$

The originating particle state for the first was originally called τ and the second θ. Both appeared in the interactions of charged strange mesons, and both were soon found to have a very similar lifetime and rest mass. The famous question thus arose: might these be the very same particle? This was known as the $\theta - \tau$ puzzle. Here is where Curie's principle appears: if θ and τ are the same, then parity symmetry is violated by Proposition 4. For whether or not parity preserves the originating particle state, it would still sometimes decay into a state with different parity, as in Figure 3. This led Lee and Yang to suggest:

> One might even say that the present $\theta - \tau$ puzzle may be taken as indication that parity conservation is violated in weak interactions. This argument is, however, not to be taken seriously because of the paucity of our present knowledge concerning the nature of the strange particles. (Lee and Yang 1956, pg.254).

Their hesitant suggestion was famously vindicated experimentally by Chien-Shiung Wu and her collaborators that same year, in an elegant experiment that was quickly repeated[14].

Curie's principle was even more directly applied in the discovery of CP-violation a few years later. A number of simple theoretical models had arisen in which the observed parity-violation was explained in a way that required CP-invariance. This requirement thus was tested by James Cronin and Val Fitch at Brookhaven, by observing a beam of neutral K-mesons or kaons. They began with a "long-lived" neutral kaon state K_L, which was known to have its phase reversed by the CP transformation;

[14]Confirming results were reported by Wu et al. (1957) and by Garwin et al. (1957).

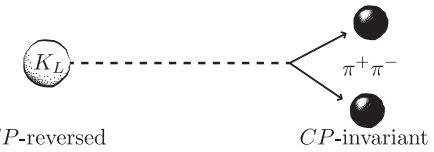


FIGURE 4. The CP-violating interaction discovered by Cronin and Fitch.

the two-pion state $\pi^+\pi^-$, on the other hand, was preserved by parity:

$$CPK_L = -K_L$$
$$CP\pi^+\pi^- = \pi^+\pi^-.$$

In a small but unmistakable number of decays, Cronin and Fitch found[15] the K_L state to decay into $\pi^+\pi^-$, as in Figure as in Figure 4. Again Curie's principle appears in the form of Proposition 4, which implies that, contrary to what the early models suggested, CP symmetry is violated.

These were two of the most important experimental discoveries of 20th century physics. Nobel prizes were awarded for each. And both were crucially underpinned by Curie's principle. In this sense, Curie was not mistaken when he suggested that "there is interest in introducing into the study of physical phenomena the symmetry arguments familiar to crystallographers" (Curie 1894)[16].

4.3. **Norton on Curie's Truism.** Norton (2014) has convincingly argued that the only true formulation of Curie's principle that does not invoke dubious causal metaphysics is a near-tautology. Namely, suppose one presumes that,

> *Determination respects symmetries:* Causes admitting symmetries are mapped to effects that admit those same symmetries.

Then Curie's claim that symmetries of the causes are symmetries of the effects is obviously true. Norton refers to this as "Curie's Lemma," pointing out:

> "there is little substance to it. It is a tautology implementing as an easy modus ponens 'A, if A then B; therefore B.' That simplicity does make precise the sense that the principle somehow has to be true." (Norton 2014, pg.6)

Let me add two comments about the little bit of substance that the truism retains, in light of what we have discussed so far.

First, establishing the truth of the premise that "Determination respects symmetries" may by itself amount to a deep result. It is analogous to the "symmetry

[15]The discovery was reported in Christenson et al. (1964).

[16]Translation from Brading and Castellani (2003, pg.311).

preservation" premise in my formulation of Curie's principle in Proposition 3, which says that given a mapping $D : C \to E$ between sets and a symmetry represented by,

(symmetry preservation) $D\sigma_c^{-1}x = \sigma_e^{-1}Dx$ for all $x \in C$.

This may be far from obvious for a given choice of causes C, effects E, and a determination relation D. The theorem formulated in section on electromagnetism establishes it for central fields in electrostatics, which is established by premises (1)-(3). But although the proof itself is straightforward, it does rely on some facts such as Stokes' theorem and the uniqueness of a compatible derivative operator that are not exactly trivial (details can be found in the Appendix).

Second, statements of the truism may have empirical significance that is non-trivial. We have seen that the discoveries of parity violation and of CP violation both involved the existence of decay modes in scattering experiments that have different symmetries from the originating states. Curie's principle establishes that such an observation is enough to tell us something interest that the laws of nature, in that there exist possible trajectories whose symmetry-transformed counterparts are not possible. In particular, the unitary evolutions corresponding to the S matrix for a weak interaction are symmetry-violating.

Curie's principle is of course still a pretty insubstantial statement in this context, in that it is still a simple fact about mappings between sets as in Proposition 3. However, this lack of substance is also a strength: the piddling amount of mathematical structure in Curie's principle assures that it is very robust. Thus, using Curie's principle to establish that the laws of nature are symmetry-violating provides evidence that is extremely resilient to theory change, even as new mathematical structures come and go[17].

5. Conclusion

Curie managed to hazard a conjecture that is of interest both when it is true and when it is false. The original hazard requires very special circumstances in order to be true. We have verified mathematically that one such circumstance is that of Curie's example, when one restricts attention to the electrostatics of central fields. However, its formulation as a general statement about electromagnetic currents and fields is false.

When we draw out the special circumstances under which Curie's hazard holds, we find a skeletal but true proposition about sets. This proposition captures what many philosophers of science have in mind when referring to "Curie's principle." Although so bare as to be nearly a triviality, formulating Curie's principle in this

[17]See Ashtekar (2014) for a more elaborate argument on this point.

way allows one to identify it among the arguments for the great symmetry-violating experiments of the mid-20th century. Viewed from this perspective, Curie's principle is indeed a simple and true statement, which managed to become one of the very fruitful symmetry principles of modern physics.

Appendix

Definitions. Let M and $\tilde{M}$ be smooth manifolds, each with a metric g_{ab} and $\tilde{g}_{ab}$, assumed here to be smooth symmetric invertible tensor fields, which are non-degenerate but not necessarily positive-definite. I adopt Penrose's abstract index notation for this discussion.

A *derivative operator* ∇ (or a 'covariant derivative' or a 'connection') maps an index a and an arbitrary tensor like λ^{bc}_{d} to another tensor, written $\nabla : (a, \lambda^{bc}_{d}) \mapsto \nabla_a \lambda^{bc}_{d}$. It is defined by the following properties, which we adopt following Malament (2012, §1.7).

(1) ∇ commutes with addition, index substitution and contraction on tensor fields.
(2) ∇ satisfies the Leibniz rule with respect to tensor multiplication.
(3) If ξ^a is a vector field and α a scalar field, then $\xi^a \nabla_a \alpha = \xi(\alpha)$. That is, $\xi^a \nabla_a \alpha$ is the "directional derivative" that ξ assigns to α.
(4) ∇ is torsion-free, in that if α is a scalar field, then $\nabla_a \nabla_b \alpha = \nabla_b \nabla_a \alpha$.

Let ∇ be a derivative operator on M, and suppose that it is *compatible* with the metric g_{ab} in that $\nabla_a g_{bc} = \mathbf{0}$. Let $\tilde{\nabla}$ similarly be a derivative operator on $\tilde{M}$ satisfying $\tilde{\nabla}_a \tilde{g}_{bc} = \mathbf{0}$. I will write $\varphi : M \to \tilde{M}$ to indicate a diffeomorphism, with pushforward φ_* and pullback φ^*.

Preserving derivatives. As a simple example, consider first the case of the derivative of a scalar field, $\nabla_a \alpha$. Every diffeomorphism φ "preserves" covariant derivatives of a scalar field, in that,

$$\varphi_*(\nabla_a \alpha) = \tilde{\nabla}_a \varphi_* \alpha. \tag{2}$$

This statement can be quickly verified: if α is any scalar field at $p \in M$ and $\tilde{\xi}^a$ is any vector at $\varphi(p) \in \tilde{M}$, then,

$$\tilde{\xi}^a \varphi_*(\nabla_a \alpha) = (\varphi^* \tilde{\xi}^a)(\nabla_a \alpha) = (\varphi^* \tilde{\xi})(\alpha) = \tilde{\xi}(\alpha \circ \varphi^{-1}) = \tilde{\xi}^a \tilde{\nabla}_a \varphi_* \alpha.$$

Equation 2 does not always hold when α is replaced with an arbitrary tensor. However, it does when we further restrict φ to be an isometry — and in fact for a slightly weaker condition. It is established by the following.

Lemma 1. *Let $\varphi : M \to \tilde{M}$ be a diffeomorphism. Then the equality,*

$$\varphi_*(\nabla_a \lambda_d^{bc}) = \tilde{\nabla}_a \varphi_* \lambda_d^{bc}$$

holds for an arbitrary tensor field like λ_d^{bc} if and only if $\tilde{\nabla}_a \varphi_* g_{ab} = \mathbf{0}$, where g_{ab} is the metric compatible with ∇. In particular the equality holds if φ is an isometry.

Proof. The 'only if' direction is trivial, since if the above equality holds for all tensors, then in particular,

$$\tilde{\nabla}_a \varphi_* g_{bc} = \varphi_*(\nabla_a g_{bc}) = \varphi_* \mathbf{0} = \mathbf{0},$$

where the penultimate equality applies compatibility, and the final equality the fact that $\varphi_* \mathbf{0} = \varphi_*(\mathbf{0} + \mathbf{0}) = \varphi_* \mathbf{0} + \varphi_* \mathbf{0}$.

For the 'if' direction, consider the mapping $\hat{\nabla}$ defined by,

$$\hat{\nabla} : (a, \lambda_d^{bc}) \mapsto \varphi^*(\tilde{\nabla}_a \varphi_* \lambda_d^{bc}).$$

where a is an index. The first step is to show that this mapping is a derivative operator. It obviously commutes with addition, index substitution and contraction because all three maps do (φ_*, φ^* and $\tilde{\nabla}_a$). It is also easy to check that it satisfies the Leibniz rule and the torsion-freeness condition. Moreover, for all vectors ξ^n and all scalar fields α, $\hat{\nabla}$ satisfies the condition that,

$$\xi^a \hat{\nabla}_a \alpha = \xi^a \varphi^*(\nabla_a \varphi_* \alpha) = \xi^a \nabla_a \varphi^* \varphi_* \alpha = \xi^a \nabla_a \alpha = \xi(\alpha),$$

where the second equality is an application of Equation 2. Therefore $\hat{\nabla}$ is a derivative operator. Note that this argument required only that φ be a diffeomorphism.

The second step is to observe that $\hat{\nabla}$ is compatible with the metric:

$$\hat{\nabla}_a g_{bc} = \varphi^* \tilde{\nabla}_a (\varphi_* g_{bc}) = \varphi^* \mathbf{0} = \mathbf{0},$$

where the second equality applies our assumption. Compatibility holds in particular when φ is an isometry, since then $\tilde{\nabla}_a(\varphi_* g_{bc}) = \tilde{\nabla}_a(\tilde{g}_{bc}) = \mathbf{0}$.

Finally, we use the fact that there is a unique derivative operator compatible with a given metric (Malament 2012, Prop. 1.9.2). So, $\hat{\nabla}$ and ∇ are the same. Therefore,

$$\nabla_a \lambda_d^{bc} = \hat{\nabla}_a \lambda_d^{bc} = \varphi^*(\tilde{\nabla}_a \varphi_* \lambda_d^{bc})$$

Pushing-forward the left and right sides with φ_*, we thus have that,

$$\varphi_*(\nabla_a \lambda_d^{bc}) = \tilde{\nabla}_a \varphi_* \lambda_d^{bc}.$$

□

As a special case of this lemma we have Proposition 1 from page 5.

Proposition 1. *If $\varphi : M \to M$ is an isometry and λ_d^{bc} an arbitrary tensor field, then $\varphi_*(\nabla_a \lambda_d^{bc}) = \nabla_a \varphi_* \lambda_d^{bc}$.*

Non-isometries. An example of a non-isometry that preserves covariant derivatives is any 'constant' conformal transformation, i.e. a conformal transformation $\varphi_* g_{ab} = \Omega^2 \tilde{g}_{ab}$ for which the conformal factor Ω is a constant scalar field, $\nabla_a \Omega = 0$. Then,

$$\tilde{\nabla}_a \varphi_* g_{ab} = \tilde{\nabla}_a (\Omega^2 \tilde{g}_{ab}) = \Omega^2 \tilde{\nabla}_a g_{ab} = \mathbf{0}.$$

Since the premises of the proposition are satisfied, this transformation φ preserves covariant derivatives.

However, these are the *only* conformal transformations that preserve covariant derivatives. If Ω is any conformal factor with non-zero covariant derivative, then applying the chain rule we have,

$$\tilde{\nabla}_a \varphi_* g_{ab} = \tilde{\nabla}_a \Omega^2 g_{ab} = g_{ab} \tilde{\nabla}_a (\Omega^2) + \Omega^2 \underbrace{\tilde{\nabla} \tilde{g}_{ab}}_{=\mathbf{0}} = 2\Omega \tilde{g}_{ab} \tilde{\nabla}_a \Omega \neq \mathbf{0}.$$

So, conformal transformations do not in general preserve covariant derivatives.

Proposition 2. *Let (M, g_{ab}) be an oriented simply-connected 3-dimensional Riemannian manifold, and let ξ^a and χ^b be two vector fields that each satisfy the Central Field assumption with respect to some (possibly different) region. If* $\mathrm{div}(\xi) = \mathrm{div}(\chi)$ *and* $\mathrm{curl}(\xi) = \mathrm{curl}(\chi)$*, then $\xi^a = \chi^a$.*

Proof. Let $\lambda^a = \xi^a - \chi^a$. We will show that $\lambda^a = \mathbf{0}$. By the linearity of the divergence and curl we have,

$$\mathrm{div}(\lambda) = \mathrm{div}(\xi) - \mathrm{div}(\chi) = 0,$$
$$\mathrm{curl}(\lambda) = \mathrm{curl}(\xi) - \mathrm{curl}(\chi) = \mathbf{0}.$$

A vanishing curl $\mathrm{curl}(\lambda) = \epsilon^{abc} \nabla_a \lambda_b$ is only possible if $\nabla_{[a} \lambda_{b]} = \mathbf{0}$, i.e. if λ_b is closed[18]. But a closed covector on a simply connected manifold is exact, meaning that it may be expressed as a gradient,

$$\lambda_a = \nabla_a \phi$$

for some scalar field ϕ (Malament 2012, Prop. 1.8.3).

Now, we have assumed that ξ^a vanishes on the boundary and outside of some region R_1, and χ^a similarly for some region R_2. Both ξ^a and χ^a thus vanish on the boundary and outside of the combined region $R = R_1 \cup R_2$, and therefore so does $\lambda^a = \xi^a - \chi^a$. That is, λ^a is a central field with respect to the region R. We thus have,

$$\int_R \lambda_a \lambda^a = \int_R (\nabla_a \phi)(\nabla^a \phi) = \int_R \nabla_a (\phi \nabla^a \phi) = \int_{\partial R} \eta_a \phi \nabla^a \phi = \int_{\partial R} \eta_a \phi \lambda^a = 0, \tag{3}$$

[18] An antisymmetric tensor $\xi_{[a} \nabla_b \lambda_{c]}$ can always be written in terms of the volume element as $\xi_{[a} \nabla_b \lambda_{c]} = k \epsilon_{abc} \epsilon^{def} \xi_{[d} \nabla_e \lambda_{f]}$ for some constant k (Malament 2012, §1.11). And a vanishing curl implies $k \epsilon_{abc} \epsilon^{def} \xi_d \nabla_e \lambda_f = \mathbf{0}$ for any arbitrary vector ξ^d. But then the total antisymmetry of ϵ_{abc} implies that $\mathbf{0} = k \epsilon_{abc} \epsilon^{def} \xi_{[d} \nabla_e \lambda_{f]} = \xi_{[d} \nabla_e \lambda_{f]}$. Since ξ^d was arbitrary, this requires $\nabla_{[e} \lambda_{f]} = \mathbf{0}$.

where second equality follows from the chain rule and the fact that $\nabla_a \nabla^a \phi = \nabla_a \lambda^a = 0$; the third equality applies Stokes' theorem (Wald 1984, Appendix B, B.2.26); and the last equality applies the assumption that $\nabla^a \phi = \lambda^a = \mathbf{0}$ on the boundary ∂R.

Finally, g_{ab} is assumed to be positive definite. Thus, $\lambda^a \lambda_a$ is strictly non-negative, so Equation 3 is only possible if $\lambda^a = \mathbf{0}$. □

References

Arfken, G. (1985). *Mathematical Methods for Physicists*, 3rd edn, San Diego: Academic Press, Inc.

Ashtekar, A. (2014). Response to Bryan Roberts: A new perspective on T violation. Forthcoming in *Studies in History and Philosophy of Modern Physics*, doi:10.1016/j.shpsb.2014.07.001.

Baez, J. and Muniain, J. P. (1994). *Gauge fields, knots and gravity*, Series on Knots and Everything Vol. 4, London: World Scientific Publishing.

Belot, G. (2003). Notes on symmetries, *in* K. Brading and E. Castellani (eds), *Symmetries in Physics: Philosophical Reflections*, Cambridge: Cambridge University Press, chapter 24, pp. 393–412.

Brading, K. and Castellani, E. (2003). *Symmetries in physics: philosophical reflections*, Cambridge: Cambridge University Press.

Brading, K. and Castellani, E. (2013). Symmetry and symmetry breaking, *in* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, spring 2013 edn.

Castellani, E. (2003). On the meaning of symmetry breaking, *in* K. Brading and E. Castellani (eds), *Symmetries in Physics: Philosophical Reflections*, Cambridge: Cambridge University Press, chapter 19, pp. 321–334.

Christenson, J. H., Cronin, J. W., Fitch, V. L. and Turlay, R. (1964). Evidence for the 2π decay of the k_2^0 meson, *Phys. Rev. Lett.* **13**(4): 138–140.

Curie, P. (1894). Sur la symétrie dans les phénomènes physique, symétrie d'un champ électrique et d'un champ magnétique, *Journal de Physique Théorique et Appliquée* **3**: 393–415.

Earman, J. (2004). Curie's Principle and spontaneous symmetry breaking, *International Studies in the Philosophy of Science* **18**(2-3): 173–198.

Garwin, R. L., Lederman, L. M. and Weinrich, M. (1957). Observations of the failure of conservation of parity and charge conjugation in meson decays: the magnetic moment of the free muon, *Physical Review* **104**(4): 1415–1417.

Ismael, J. (1997). Curie's Principle, *Synthese* **110**(2): 167–190.

Lee, T. D. and Yang, C.-N. (1956). Question of Parity Conservation in Weak Interactions, *Physical Review* **104**(1): 254–258.

Malament, D. B. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*, Chicago: University of Chicago Press.

Mittelstaedt, P. and Weingartner, P. A. (2005). *Laws of Nature*, Springer-Verlag Berlin Heidelberg.

Norton, J. D. (2014). Curie's Truism. Unpublished Manuscript, `http://philsci-archive.pitt.edu/10926/`.

Roberts, B. W. (2013a). The Simple Failure of Curie's Principle, *Philosophy of Science* **80**(4): 579–592.

Roberts, B. W. (2013b). Three merry roads to T-violation. `http://philsci-archive.pitt.edu/9632/`.

Wald, R. M. (1984). *General Relativity*, Chicago: University of Chicago Press.

Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. and Hudson, R. P. (1957). Experimental test of parity conservation in beta decay, *Physical Review* **104**(4): 1413–1415.

Is de Broglie-Bohm Theory Specially Equipped to Recover Classical Behavior?

Joshua Rosaler

Abstract

Supporters of the de Broglie-Bohm (dBB) interpretation of quantum theory argue that because the theory, like classical mechanics, concerns the motions of point particles in 3D space, it is specially suited to recover classical behavior. I offer a novel account of classicality in dBB theory, if only to show that such an account falls out almost trivially from results developed in the context of decoherence theory. I then argue that this undermines any special claim that dBB theory is purported to have on the unification of the quantum and classical realms.

1 Introduction

Several advocates of the de Broglie-Bohm (dBB) interpretation of quantum theory hold that because, like classical mechanics, it concerns the motions of point particles in 3D space, it is specially suited to recover classical behavior. [1] [2] They note that in dBB theory, we can ask simply: under what circumstances do the additional particle configurations posited by the theory follow approximately Newtonian trajectories? Moreover, the equations of motion for the additional "hidden" variables in dBB theory take the form of classical equations of motion, but with an additional "quantum potential" or "quantum force" term that produces deviations of the trajectories from classicality. On this basis, a number of authors have suggested that dBB theory furnishes special tools to recover classical behavior that other interpretations do not, via the requirement that the quantum potential or force be negligibly small (Allori et al., 2002), (Bohm and Hiley, 1995).

[1] By "classical," I mean having sharply defined values for properties such as position and momentum *and* conforming approximately to Newtonian equations of motion.

[2] For brevity, I refer to the de Broglie-Bohm interpretation of quantum theory as "dBB theory."

Here, I provide an alternative to existing accounts of classicality in dBB theory, if only to show that such an account falls out almost trivially from results developed outside the context of dBB theory, in the literature on decoherence. Formal tools specific to dBB theory, such as the quantum potential or quantum force, turn out to be neither necessary nor helpful to the analysis. Here, I regard decoherence theory as an interpretation-neutral body of results that follow when both a system and its environment (including any observers, measuring apparati, and residual microscopic degrees of freedom) are modeled as a closed system that always obeys the unitary Schrodinger evolution. On my account of classicality in dBB theory, the Bohmian configuration evolves classically only because it succeeds in tracking one among the many branches of the total quantum state defined by decoherence, and the approximate classicality of such a branch in turn is a consequence of the unitary Schrodinger evolution, on which dBB theory has no special claim.

Since the dBB account of classicality is entirely parasitic on the branching structure defined by decoherence, the claim that dBB theory is uniquely suited to recover classical behavior *must already presuppose what is at issue* in debates about the interpretation of quantum mechanics: namely, that it provides the one true account of how one among the many potentialities/branches contained in a decohered quantum superposition gets selected as "the outcome" (necessarily, in accordance with Born Rule). Insofar as other interpretations furnish viable resolutions to this issue, they should be able to make similar use of decoherence-based results to recover classical behavior on their own terms. For example, (Wallace, 2012), Ch. 3 provides a decoherence-based account of macroscopic classical behavior on the Everett Interpretation. At most, the fact that dBB theory can recover classical behavior on its own terms can be regarded as an internal test of the theory's adequacy; it should not be regarded as a point in favor of dBB theory over other interpretations.

In Section 2, I review the major existing lines of approach to modelling classical behavior in dBB theory and highlight a number of gaps in these accounts. In Section 3, I explain the basic mechanism whereby environmental decoherence helps to recover macroscopic classical behavior in dBB theory. In Section 4, I summarize the analysis of macroscopic classical behavior furnished by decoherence theory on the bare formalism of quantum mechanics (that is, Schrodinger evolution without collapse) . In Section 5, I show how the dBB account of classical behavior follows straightforwardly from the decoherence-based analysis of the previous section.

2 Existing Accounts of Classicality in dBB Theory

Bohm's theory posits that the state of any closed system is given by a wave function $\Psi(X,t)$ that always evolves according to the Schrodinger equation (with the same Hamiltonian as in conventional quantum mechanics), and a configuration q that evolves according to the equation $\frac{dq}{dt} = \frac{\nabla S(X,t)|_q}{m}$, where S is the phase of the wave function; it also posits that, at some initial time t_0, our knowledge of the particle configuration is characterized by the probability distribution $|\Psi(X,t_0)|^2$ (Bohm, 1952a), (Bohm, 1952b). Efforts to model classical behavior in Bohm's theory fall into two broad categories: 1) what I call "quantum potential" approaches, which rely on the vanishing of the quantum potential and/or force, and in which the Bohmian configuration occupies center stage; 2) what I call "narrow wave packet" approaches, which take an analysis of the wave function as their starting point and treat the Bohmian configuration as being in some sense simply "along for the ride."

2.1 Quantum Potential Approaches

If one plugs the polar decomposition of the wave function, $\Psi(X,t) = R(X,t)e^{i\frac{S(X,t)}{\hbar}}$, into Schrodinger's equation, one obtains the following relations as the real and imaginary parts, respectively, of the resulting relation:

$$\frac{\partial S}{\partial t} = \frac{(\nabla S)^2}{2M} + V - \frac{\hbar^2}{2M}\frac{\nabla^2 R}{R} \tag{1}$$

$$\frac{\partial R^2}{\partial t} + \nabla \cdot (\frac{\nabla S}{M} R^2) = 0. \tag{2}$$

The first equation takes the form of the classical Hamilton-Jacobi equation, except for the additional term $Q \equiv -\frac{\hbar^2}{2M}\frac{\nabla^2 R}{R}$, known as the "quantum potential." The second equation takes the form of a continuity equation for the probability distribution $\rho \equiv R^2$. Together with the evolution equation for the dBB configuration, $\frac{dq}{dt} = \frac{\nabla S(X,t)|_q}{M}$, the first of these equations implies that this configuration obeys the relation,

$$M\frac{d^2 q}{dt^2} = -\nabla V|_q - \nabla Q|_q, \tag{3}$$

which mathematically resembles Newton's Second Law, but with an additional "quantum force" term $-\nabla Q$. When $\nabla Q \approx 0$, q follows an approximately Newtonian trajectory; some authors have also suggested $Q \approx 0$ as a requirement for classicality, though the sufficiency of this condition for classicality presupposes implicitly that $\nabla Q \approx 0$ is also satisfied Many supporters of dBB theory feel that the quantum potential/force's being approximately equal to zero furnishes a simple, transparent condition for classicality, and moreover, one that is unique to dBB theory. Bohm and Hiley, Holland, and Allori, Durr, Goldstein and Zanghi are among the authors who offer analyses of classicality rooted in this supposition (Bohm, 1952a), (Bohm and Hiley, 1995), (Holland, 1995), (Allori et al., 2002), (Durr and Teufel, 2009). Some of these authors also suggest ways of generalizing this approach to incorporate environmental decoherence, but do not explain in detail how the conditions $Q \approx 0$, $\nabla Q \approx 0$ are to be extended to the case where the system is open, or why this would be useful given that the results (1) and (3) were derived under the now-abandoned assumption that the system is closed and in a pure state.

2.2 Narrow Wave Packet Approaches

A second approach that is sometimes adopted in the effort to model classical behavior in dBB theory makes use of Ehrenfest's Theorem,

$$M\frac{d\langle\hat{P}\rangle}{dt} = -\langle\frac{\partial\hat{V}}{\partial X}\rangle, \tag{4}$$

which applies generally to wave functions evolving under the Schrodinger equation with a Hamiltonian of the form $\hat{H} = \frac{\hat{P}^2}{2M} + V(\hat{X})$. The crucial point in this approach is that for wave packets narrowly peaked in position (narrowly, that is, relative to some characteristic length scale on which the potential V changes), one has approximately that

$$M\frac{d\langle\hat{P}\rangle}{dt} \approx -\frac{\partial V(\langle\hat{X}\rangle))}{\partial\langle\hat{X}\rangle}, \tag{5}$$

which entails that the expectation value of position $\langle\hat{X}\rangle$ evolves approximately along a Newtonian trajectory. In dBB theory, the property of equivariance, whereby the Born Rule probability distribution $|\Psi(X,t)|^2$ is preserved by the flow of the configuration variables, ensures that the Bohmian

trajectory will follow the wave packet and therefore traverse the same Newtonian trajectory as the expectation value $\langle \hat{X} \rangle$. The primary advocate of the narrow wave packet approach in the literature has been Bowman, who has also actively criticized approaches based on the quantum potential and quantum force (Bowman, 2005). Bowman notes that the narrow wave packet approach, as applied to isolated systems, does not explain why the state of S should be a narrow wave packet to begin with; however, he argues, correctly on my view, that this can be corrected by incorporating environmental decoherence into the analysis. However, Bowman's account does not recognize the need, not just for decoherence, but for a special type of decoherence that ensures disjointness of the branches of the total quantum state in the combined configuration space of the system and environment. Moreover, Bowman confines his attention to the *reduced* dynamics of the open system whose classicality we seek to model (say, the center of mass of the moon) and does not consider the structure of the overall pure state of the system and environment; for reasons that will become clear in the next section, this restriction obscures the true mechanism whereby the Bohmian configuration of the system comes to be guided by just one of the narrow wave packets present in the overall superposition.

3 The Role of the Environment

To illustrate the role of the environment in recovering classicality from dBB theory, and why it is not sufficient to model a macroscopic system like the moon's center of mass as an isolated system, it is instructive to consider a simple example. Let S be the center of mass of some macroscopic body, and assume to begin with that the system is always closed and in a pure state. Let the pure state be a superposition of narrow wave packets with opposite average momenta, initially separated across a macroscopically large distance in space; in addition, let the time evolution of the state be such that the trajectories of the packets - for simplicity, assume they are straight lines and that the system is free - overlap at some point in S's configuration space and then pass through each other (so that the trajectories of the two packets over time make the shape of an "X"). Then the quantum state at every time takes the form,

$$|\Psi\rangle = \frac{1}{\sqrt{2}}[|q_1, p\rangle + |q_2, -p\rangle], \tag{6}$$

where $|q, p\rangle$ designates the quantum state of a wave packet simultaneously peaked about postion q and momentum p (to within the restrictions of the uncertainty principle) and q_1 and q_2 change with time in the manner specified. Assuming the mass M to be macroscopically large, we can neglect spreading of these wave packets. [3] Now consider an ensemble of initial Bohmian configurations associated with this pure state. Those trajectories associated with initial conditions in the first packet initially will follow the classical straight-line trajectory of that packet, and likewise for the second packet. However, Bohmian trajectories of a closed pure-state system cannot cross, so when the packets overlap in configuration space, trajectories initially associated with one wave packet will exit the overlap region in the wave packet in which they did *not* begin, rather than proceeding in a straight line with their original wave packet. Thus, the trajectories will exhibit highly non-classical kink as a consequence of the overlap. In dBB theory, such non-classicalities in the Bohmian trajectory are a generic consequence of wave packets overlapping in configuration space, even in cases where the mass M is macroscopically large ($M \gtrsim 1kg.$) and wave packet spreading can be neglected; also, they are not restricted to the simple case of a free particle discussed here.

Let us now abandon the assumption that the system S is isolated and allow it to interact with and become entangled with its environment E. Now it is the combined system SE, rather than S, that is closed and in a pure state, though it is still S's classicality that we wish to recover. Let us assume that at every time the wave function of the closed system SE consisting of the center of mass and its environment (which may consist of photons, neutrinos, or other particles of matter) takes the form

$$|\Psi\rangle = \frac{1}{\sqrt{2}}[|q_1, p\rangle \otimes |\phi_1\rangle + |q_2, -p\rangle \otimes |\phi_2\rangle], \tag{7}$$

for some states $|\phi_1\rangle$ and $|\phi_2\rangle$ in E's Hilbert space $\mathcal{H}_E$, where $|q_1, p\rangle$ and $|q_2, p\rangle$ follow the same trajectories through S's configuration space as in the

[3] Using the formula $\sigma(t) = \sqrt{\sigma_0^2 + (\frac{\hbar t}{M\sigma_0})^2}$ for the time dependence of the width of an initial Gaussian under free evolution, one can show that for a free particle of mass $M \sim 1kg.$, the time it takes for a wave packet initially localized on the scale of an Angstrom to spread to a centimeter is longer than the age of the universe.

isolated case just considered. Moreover, assume that $|\phi_1\rangle$ and $|\phi_2\rangle$ have disjoint supports in E's configuration space $\mathbb{Q}_E$ - that is, that they are "superorthogonal": [4]

$$\langle\phi_1|y\rangle\langle y|\phi_2\rangle \approx 0 \text{ for all } y \in \mathbb{Q}_E \tag{8}$$

where $|y\rangle$ is a position (or more accurately, configuration) eigenstate of the environment. Because of the disjointness of the supports of $|\phi_1\rangle$ and $|\phi_2\rangle$ in $\mathbb{Q}_E$, the packets $|q_1, p\rangle\otimes|\phi_1\rangle$ and $|q_2, -p\rangle\otimes|\phi_2\rangle$ will remain disjoint in the total configuration space $\mathbb{Q}_{SE}$, even when $|q_1, p\rangle$ and $|q_2, -p\rangle$ overlap in $\mathbb{Q}_S$. The non-overlap condition for Bohmian trajectories applies only to trajectories in $\mathbb{Q}_{SE}$ since the state is pure only relative to SE, so the configuration $q_{SE} = (Q_S, q_E)$ of the whole system will forever remain in one wave packet or the other - say, the first one. As a consequence, the configuration Q_S associated with S always follows the classical trajectory of the wave packet $|q_1, p\rangle$ in $\mathbb{Q}_S$. There is no "kink" as in the isolated case. If E contains on the order of 10^{23} microscopic degrees of freedom, as it typically will, we can expect the relation (8) to hold irreversibly, and for this reason can effectively ignore the second wave packet since it will have no influence on the motion of the total configuration. Moreover, the configuration q_E of the environment will be irreversibly correlated to the wave packet $|q_1, p\rangle$, since it is bound lie in the support of $\langle y|\phi_1\rangle$ (if q_{SE} had started in the second packet, q_E would instead be in the disjoint region associated with the support of $\langle y|\phi_2\rangle$ and be correlated with the wave packet $|q_2, -p\rangle$).

4 Decoherence-Based Models of Classical Behavior

In this section, I describe the evolution of the pure state of a closed system SE consisting of the center of mass S of some macroscopic body and its environment E, on the assumption that this evolution is always governed by the Schrodinger equation. This analysis draws most directly from (Joos

[4]By "support" of a configuration space function, I mean the region of configuration space in which the function's value is not negligibly small - so, greater than some arbitrarily chosen small ϵ. Note that superorthogonality implies orthogonality but is not implied by it; the term "superorthogonal" can be traced back to (Bohm and Hiley, 1995) and (Maroney, 2005) .

et al., 2003), Ch.'s 3 and 5, (Hartle, 2011), and (Wallace, 2012), Ch.3. In the next section, I show that an account of classicality at the level of the Bohmian configuration follows straightforwardly from this analysis.

The quantum description of the closed system in question takes as its Hilbert space $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_E$, the tensor product of the Hilbert space $\mathcal{H}_S$ associated with the center of mass of the body in question and the Hilbert space $\mathcal{H}_E$ associated with the residual microscopic degrees of freedom in the environment (which includes degrees of freedom both internal and external to the body in question). The dynamics with respect to this set of variables are given by a Schrodinger equation of the form

$$i\hbar\frac{\partial|\Psi\rangle}{\partial t} = \left(\hat{H}_S \otimes \hat{I}_E + \hat{I}_S \otimes \hat{H}_E + \hat{H}_I\right)|\Psi\rangle, \tag{9}$$

where $|\Psi\rangle \in \mathcal{H}_S \otimes \mathcal{H}_E$, $\hat{I}_E$ is the identity operator on $\mathcal{H}_E$, $\hat{I}_S$ the identity operator on $\mathcal{H}_S$, and $\hat{H}_I$ is an interaction Hamiltonian acting on $\mathcal{H}_S \otimes \mathcal{H}_E$. In the models of interest here, $\hat{H}_S = \frac{\hat{P}^2}{2M} + V(\hat{X})$, and $\hat{H}_I$ is a function of only of center-of-mass position $\hat{X}$ and the positions of environmental particles, represented collectively by $\hat{y}$. At a more coarse-grained level, we can examine the evolution of the reduced density matrix $\hat{\rho}_S \equiv Tr_E|\Psi\rangle\langle\Psi|$ of S. For a wide variety of models, in which environmental decoherence is significant but dissipative effects can be ignored, the evolution of $\hat{\rho}_S$ is governed by the equation,

$$i\hbar\frac{\partial\hat{\rho}_S}{\partial t} = \big[\hat{H}_S, \hat{\rho}_S\big] - i\Lambda\big[\hat{X}, \big[\hat{X}, \hat{\rho}_S\big]\big], \tag{10}$$

where the first term generates unitary evolution prescribed by $\hat{H}_S$ and the second represents the effect of decoherence from the environment; the second term suppresses the off-diagonal elements of $\langle X'|\hat{\rho}_S|X\rangle$ throughout its evolution, and Λ is a constant derived from the parameters in the closed system Hamiltonian in (9) (This is an important special case of the well-known Caldeira-Leggett equation; for further discussion and derivation of this equation, see (Joos et al., 2003) and (Schlosshauer, 2008)). From (10), one can show that $M\frac{d\langle\hat{P}\rangle}{dt} = -\langle\frac{\partial\hat{V}}{\partial X}\rangle$, where $\langle\hat{O}\rangle \equiv Tr[\hat{\rho}_S\hat{O}]$ for any Hermitian operator $\hat{O}$ on $\mathcal{H}_S$; this constitutes a generalization of Ehrenfest's Theorem to open, decohering quantum systems (Joos et al., 2003). By analogy with the case of closed systems, one can show that when the width of the distribution $\rho_S(X) \equiv \langle X|\hat{\rho}_S|X\rangle$, known as the ensemble width of S, is narrow

by comparison with the characteristic length scales on which V varies, we have $M\frac{d\langle\hat{P}\rangle}{dt} \approx -\frac{\partial V(\langle\hat{X}\rangle)}{\partial\langle\hat{X}\rangle}$, which entails that the expectation value of position $\langle\hat{X}\rangle = Tr_S(\hat{\rho}_S\hat{X})$ follows an approximately Newtonian trajectory as long as the width of the distribution $\rho_S(X)$, also known as the ensemble width of $\hat{\rho}_S$, remains narrowly peaked relative to the characteristic length scales on which V varies. The timescales on which the ensemble width of an initially narrow $\rho_S(X)$ remains narrowly peaked will depend both on the value of the mass M and on the strength of chaotic effects in the Hamiltonian $\hat{H}_S$ (for further discussion of the role of chaos in wave packet spreading in open systems, see (Zurek and Paz, 1995)).

Let us now consider what constraints this analysis of $\hat{\rho}_S$ places on the the evolution of the pure state $|\Psi\rangle$ of the total system SE, recalling that $\hat{\rho}_S \equiv Tr_E|\Psi\rangle\langle\Psi|$. The decoherent or consistent histories formalism will prove especially useful for this purpose. [5] Consider a partition $\{\Sigma_\alpha\}$ of the classical phase space associated with the system S such that the cells Σ_α all have equal phase space volume. Using this partition, we can define the positive operator-valued measure (POVM) given by the operators $\hat{\Pi}_\alpha \equiv \int_{\Sigma_\alpha} dz\ |z\rangle\langle z|$, where $z \equiv (q,p)$ is a notational shorthand for a point in phase space, and $|z\rangle$ is a minimum-uncertainty coherent state centered on the phase space point z. [6] [7] If the cells Σ_α are significantly larger than the volume in phase space over which coherent states have strong support (i.e., $\hbar$), then the operators $\hat{\Pi}_\alpha$ constitute an approximate PVM since in this case $\hat{\Pi}_\alpha\hat{\Pi}_\beta \approx \delta_{\alpha\beta}\hat{\Pi}_\alpha$. We can extend this approximate PVM on $\mathcal{H}_S$ to an approximate PVM $\{\hat{P}_\alpha\}$ on $\mathcal{H}_S \otimes \mathcal{H}_E$ by defining $\hat{P}_\alpha = \hat{\Pi}_\alpha \otimes \hat{I}_E$. Inserting factors of the identity $\hat{I}_{SE} = \sum_{\alpha_i} \hat{P}_{\alpha_i}$ at regular time intervals of the unitary evolution, we can then write the state evolution at successive time intervals $N\Delta t$ as follows:

[5] For an introduction to the decoherent histories formalism, see for example (Gell-Mann and Hartle, 1993), (Griffiths, 1984), (Halliwell, 1995).

[6] For my purposes, it is sufficient for the reader to think of a coherent state state simply as a Gaussian wave packet narrowly peaked both in position and momentum.

[7] A positive-operator-valued measure (POVM) on $\mathcal{H}$ is a set $\{\hat{\Pi}_\alpha\}$ of positive operators such that $\sum_\alpha \hat{\Pi}_\alpha = \hat{I}$; recall that an operator $\hat{O}$ is positive if it is self-adjoint and $\langle\Psi|\hat{O}|\Psi\rangle \geq 0$ for every $|\Psi\rangle \in \mathcal{H}$. A projection-valued measure (PVM) $\{\hat{P}_\alpha\}$ on Hilbert space $\mathcal{H}$ is a POVM such that $\hat{P}_\alpha\hat{P}_\beta = \delta_{\alpha\beta}\hat{P}_\alpha$ (no summation over repeated indices).

$$|\Psi(N\Delta t)\rangle = e^{-\frac{i}{\hbar}\hat{H}N\Delta t}|\Psi_0\rangle \tag{11}$$

$$= (\sum_{\alpha_N} \hat{P}_{\alpha_N}) e^{-\frac{i}{\hbar}\hat{H}\Delta t} (\sum_{\alpha_{N-1}} \hat{P}_{\alpha_{N-1}}) ... (\sum_{\alpha_1} \hat{P}_{\alpha_1}) e^{-\frac{i}{\hbar}\hat{H}\Delta t} (\sum_{\alpha_0} \hat{P}_{\alpha_0}) |\Psi_0\rangle \tag{12}$$

$$= \sum_{\alpha_0,...,\alpha_N} \hat{C}_{\alpha_0,...,\alpha_N} |\Psi_0\rangle \tag{13}$$

where the components $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$ are defined by

$$\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle \equiv \hat{P}_{\alpha_N} e^{-\frac{i}{\hbar}\hat{H}\Delta t} \hat{P}_{\alpha_{N-1}} ... \hat{P}_{\alpha_1} e^{-\frac{i}{\hbar}\hat{H}\Delta t} \hat{P}_{\alpha_0}, |\Psi_0\rangle. \tag{14}$$

The reason for using this particular approximate PVM will be made clear shortly. Each component $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$, corresponds to a particular "history" or sequence $(\Sigma_{\alpha_0}, ..., \Sigma_{\alpha_N})$ of regions through phase space. Let us examine in more detail the structure of one of these components. Using the definition of the operators $\hat{P}_{\alpha_i}$ we can write,

$$\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle = \int_{\Sigma_{i_0}} ... \int_{\Sigma_{i_N}} dz_0 ... dz_N \ |z_N\rangle \otimes |\tilde{\phi}(z_0, ..., z_N)\rangle \tag{15}$$

$$= \int_{\Sigma_{i_0}} ... \int_{\Sigma_{i_N}} dz_1 ... dz_N \ w(z_0, ..., z_N) \ |z_N\rangle \otimes |\phi(z_0, ..., z_N)\rangle \tag{16}$$

with $|\tilde{\phi}(z_0, ..., z_N)\rangle \equiv \sum_i |e_i\rangle\langle z_N, e_i|\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle \in \mathcal{H}_E$ for $\{|e_i\rangle\}$ any basis of $\mathcal{H}_E$, $w(z_0, ..., z_N) \equiv \sqrt{\langle\tilde{\phi}(z_0, ..., z_N)|\tilde{\phi}(z_0, ..., z_N)\rangle}$, and $|\phi(z_0, ..., z_N)\rangle \equiv \frac{|\tilde{\phi}(z_0,...,z_N)\rangle}{w(z_0,...,z_N)}$. As Zurek has shown, the coherent states $|z\rangle$ for systems like S are special in that under fairly generic conditions, they become entangled with the environment only on much longer timescales than other states in $\mathcal{H}_S$; he calls such states "pointer states" (Zurek et al., 1993). As Wallace demonstrates in detail in (Wallace, 2012), Ch.3, continuous monitoring of the center of mass position by the environment (usually via scattering of photons, air molecules, etc. by the center of mass) enforces the relation:

$$\langle\phi(z'_0, ..., z'_N)|\phi(z_0, ..., z_N)\rangle \approx 0 \tag{17}$$

for z_i and z'_i differing by more than the width of a coherent wave packet, for any $0 \leq i \leq N$. From (15) and (17) it follows immediately that

$$\langle\Psi_0|\hat{C}^{\dagger}_{\alpha'_0,...,\alpha'_N}\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle \approx 0 \tag{18}$$

if $\alpha_i \neq \alpha_i$ for any $0 \leq i \leq N$. When this condition holds, each component $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$ of the total superposition is said to constitute a "branch" of the quantum state, or simply a branch state (note that, as written, they are not normalized). Thus, we can see the reason for the choice of the approximate coherent state PVM: the histories defined in terms of this PVM are mutually decoherent, which follows as a consequence of the fact that the states $|z\rangle$ are generically pointer states for systems like S. In turn, satisfaction of (18) for each N ensures that the only allowable transitions from branch states $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$ at an earlier time to branch states $\hat{C}_{\beta_0,...,\beta_M}|\Psi_0\rangle$ at a later time (with $N < M$), are those for which $(\beta_0, ..., \beta_N) = (\alpha_0, ..., \alpha_N)$ - that is, such that the history associated with the earlier state is an initial segment of the history associated with the later state. This is part of what is meant when decoherence is said to generate a branching structure for the quantum state.

As a consequence of the open systems version of Ehrenfest's Theorem, on time scales where ensemble spreading can be ignored, $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle \approx 0$ for all histories $(\alpha_0, ..., \alpha_N)$ that are not approximately classical. Thus, we can restrict the sum (11) to the subset $\mathbb{H}_c$ of histories that are approximately classical:

$$\boxed{|\Psi(N\Delta t)\rangle \approx \sum_{(\alpha_0,...,\alpha_N)\in\mathbb{H}_c} \hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle} \tag{19}$$

From this we can see that *relative to a single branch* $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$, the mean values of S's position and momentum at each time step $i\Delta t$ (with $1 \leq i \leq N$) lie along an approximately classical trajectory, and the ensemble distributions in position and momentum relative to this branch remain tightly peaked around these values. [8] Moreover, it follows from (17) that the reduced density matrix of E relative to branch $\alpha \equiv (\alpha_0, ..., \alpha_N)$, $\hat{\rho}^{\alpha}_E \equiv \frac{1}{|w(\alpha)|^2} Tr_S[\hat{C}_{\alpha}|\Psi_0\rangle\langle\Psi_0|\hat{C}^{\dagger}_{\alpha}]$, exhibits a strong correlation to this trajectory in that it is orthogonal to the reduced density matrix $\hat{\rho}^{\alpha'}_E$ associated with any other trajectory/branch α' - i.e., $Tr_E(\hat{\rho}^{\alpha}_E\hat{\rho}^{\alpha'}_E) \approx \delta_{\alpha\alpha'}$.

[8] Relative to the branch $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$, the expectaiton values of position and momentum at times i earlier than N are given respectively by $\frac{1}{|w(\alpha_0,...,\alpha_i)|^2}\langle\Psi_0|\hat{C}^{\dagger}_{\alpha_0,...,\alpha_i}(\hat{X} \otimes \hat{I}_E)\hat{C}_{\alpha_0,...,\alpha_i}|\Psi_0\rangle$ and $\frac{1}{|w(\alpha_0,...,\alpha_i)|^2}\langle\Psi_0|\hat{C}^{\dagger}_{\alpha_0,...,\alpha_i}(\hat{P} \otimes \hat{I}_E)\hat{C}_{\alpha_0,...,\alpha_i}|\Psi_0\rangle$.

5 The dBB Model of Macroscopic Newtonian Systems

Since the wave function in dBB theory obeys the same Schrodinger dynamics as was assumed in the analysis of the previous section, the quantum state in the corresponding dBB model also takes the form (19). However, we saw in Section (8) that classicality in Bohm's theory requires not just the orthogonality of environmental states associated with different branches, represented in (17), but the stronger condition of superorthogonality, which ensures disjointness of these states in $\mathbb{Q}_E$:

$$\langle \phi(z'_0, ..., z'_N)|y\rangle\langle y|\phi(z_0, ..., z_N)\rangle \approx 0 \ \forall y \in \mathbb{Q}_E, \tag{20}$$

for z_i and z'_i sufficiently different for any $0 \leq i \leq N$. Typically, the unitary Schrodinger evolution will also enforce this stronger condition. It follows immediately from (20) that the branch states associated with different histories are disjoint in the full configuration space $\mathbb{Q}_{SE}$:

$$\langle \Psi_0|\hat{C}^{\dagger}_{\alpha'_0,...,\alpha'_N}|X, y\rangle\langle X, y|\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle \approx 0 \ \forall (X, y) \in \mathbb{Q}_{SE} \tag{21}$$

if $\alpha_i \neq \alpha'_i$ for any $0 \leq i \leq N$. As a consequence of this disjointness, the Bohmian configuration q_{SE} will lie in the support of just one branch $\hat{C}_{\beta_0,...,\beta_N}|\Psi_0\rangle$, and the influence of all other branches on its evolution, and all future sub-branches of those other branches, can be neglected.

Let us now examine what this implies about the evolution of the system configuration Q_S and the environmental configuration q_E. Let $SE_{\beta_0,...,\beta_N}$ designate the support of $\hat{C}_{\beta_0,...,\beta_N}|\Psi_0\rangle$ in $\mathbb{Q}_{SE}$. This region will be contained in the region $S_{\beta_0,...,\beta_N} \times E_{\beta_0,...,\beta_N}$, the direct product of the regions in which the marginal distributions over $\mathbb{Q}_S$ and $\mathbb{Q}_E$ have support, with $S_{\beta_0,...,\beta_N} \equiv \text{supp}\big[\int dy|\langle X, y|\hat{C}_{\beta_0,...,\beta_N}|\Psi_0\rangle|^2\big] \subset \mathbb{Q}_S$ and $E_{\beta_0,...,\beta_N} \equiv \text{supp}\big[\int dX|\langle X, y|\hat{C}_{\beta_0,...,\beta_N}|\Psi_0\rangle|^2\big] \subset \mathbb{Q}_E$. Now the region $S_{\beta_0,...,\beta_N}$ should roughly coincide with the range of positions associated with the phase space region Σ_{β_N}; thus, $S_{\beta_0,...,\beta_N}$ for each N should lie close to the Newtonian configuration space trajectory $X_{cl}(N\Delta t)$ associated with the sequence $(\Sigma_{\beta_0}, ..., \Sigma_{\beta_N})$. Moreover, because of (20), the regions $E_{\alpha_0,...,\alpha_N}$ corresponding to the environmental support of each distinct branch will be disjoint, so that

$$E_{\alpha'_0,...,\alpha'_N} \cap E_{\alpha_0,...,\alpha_N} = \emptyset \tag{22}$$

if $\alpha_i \neq \alpha'_i$ for any $0 \leq i \leq N$.

Since $q_{SE} = (Q_S, q_E)$, and $q_{SE} \in SE_{\beta_0,...,\beta_N} \subset S_{\beta_0,...,\beta_N} \times E_{\beta_0,...,\beta_N}$, it follows that $Q_S \in S_{\beta_0,...,\beta_N}$ and $q_E \in E_{\beta_0,...,\beta_N}$. Thus, the Bohmian configuration Q_S of the system S follows an approximately Newtonian trajectory $X_{cl}(N\Delta t)$ near to that associated with the sequence $(\Sigma_{\beta_0}, ..., \Sigma_{\beta_N})$, while the configuration of the environment E becomes correlated to this trajectory and thereby serves as a record of it. So, at last, we have that

$$\boxed{|Q_S(N\Delta t) - X_{cl}(N\Delta t)| < \delta,} \tag{23a}$$

$$\boxed{q_E(N\Delta t) \in E_{\beta_0,...,\beta_N}} \tag{23b}$$

for all N such that $N\Delta t$ is less than the time when ensemble spreading of $\rho_S(X)$ becomes appreciable, and for δ some suitably chosen small margin of error. Moreover, if $q_{SE} = (Q_S, q_E)$ lies in the support of a single branch $\hat{C}_{\alpha_0,...,\alpha_N}|\Psi_0\rangle$, at later times it may be found in the support only of branches $\hat{C}_{\beta_0,...,\beta_M}|\Psi_0\rangle$ such that $(\alpha_0, ..., \alpha_N)$ are the first N indices in $(\beta_0, ..., \beta_M)$, where $N < M$. If $M\Delta t$ is less than the timescale on which ensemble spreading of $\rho_S(X)$ becomes appreciable, $(\beta_0, ..., \beta_M)$ will represent the continuation up to $M\Delta t$ of the classical trajectory approximated by $(\alpha_0, ..., \alpha_N)$. This follows from (21) and the equivariance of the Bohmian configuration's dynamics.

6 Conclusions

The analysis of classicality advanced in the previous section extends the effective collapse mechanism originally developed by (Bohm, 1952b), as applied to the context of a laboratory quantum measurement, to the context of a classically evolving macroscopic body interacting with some environment. In both cases, decoherence renders the total state a superposition of disjoint packets, so that the configuration comes to be guided by only one of these packets. [9]

[9]Some have suggested that dBB theory does not require decoherence to solve the measurement problem because the theory is already about objects with determinate positions and momenta. There are two problems with this line of thinking. First, it presupposes that decoherence is somehow optional in dBB theory. It isn't: decoherence is a generic consequence of unitary evolution, and thus happens in dBB theory whether or not the empirical adequacy of dBB theory requires it. Second, the empirical adequacy of dBB theory *does* rely on decoherence, since as Bohm showed in (Bohm, 1952b), decoherence

Given this analysis, the position that dBB theory is specially equipped to recover classical behavior must presuppose the very point that is at issue in debates about the measurement problem: namely, that the Bohmian mechanism for effectively collapsing a decohered superposition onto a single component is the true mechanism employed in nature. Because advocates of other interpretations provide their own mechanisms for the collapse or effective collapse of a decohered superposition, those who do not already submit to the dBB interpretation are unlikely to be impressed by its account of classical behavior. In particular, advocates of the Everett interpretation are likely to regard the analysis given in Section 5, concerning the evolution of the Bohmian configuration, as utterly superfluous to a quantum description of classical behavior since they regard the structure of a unitarily evolving quantum state as sufficient for this purpose; see (Brown and Wallace, 2005). However, many are also hesitant to accept that the structure associated with a unitarily evolving quantum state on its own is sufficient to save the appearances, not least because this supposition entails the existence of a vast, ever-growing proliferation of worlds associated with the different branches of the quantum state. Barring the objection that the Bohmian configuration is superfluous, the fact that dBB theory is able to support an account of classical behavior on its own terms should at least provide some reassurance of its continuing viability in the nonrelativistic domain. However, it should not be counted as an advantage of dBB theory over other interpretations.

Acknowledgments: Thanks to David Wallace, Simon Saunders, Harvey Brown, Christopher Timpson and Jeremy Butterfield for comments on earlier drafts of this work, and to Cian Dorr for helpful discussions of de Broglie-Bohm theory. Thanks also to audiences in Oxford, Sussex and Vallico Soto, Tuscany. This work was supported by the University of Oxford Clarendon fund and the University of Pittsburgh's Center for Philosophy of Science.

References

V. Allori, D. Dürr, S. Goldstein, and N. Zanghí. Seven steps towards the classical world. *Journal of Optics B: Quantum and Semiclassical Optics*, 4(4):S482, 2002.

serves as the lynchpin for the theory's effective collapse mechanism; without decoherence, the apparatus configuration in a measurement does not become irreversibly correlated to the measured system, nor does the theory recover the Born Rule for measurements of non-position observables.

D. Bohm. A suggested interpretation of the quantum theory in terms of 'hidden' variables. i. *Physical Review*, 85(2):166, 1952a.

D. Bohm. A suggested interpretation of the quantum theory in terms of 'hidden' variables. ii. *Physical Review*, 85(2):180, 1952b.

D. Bohm and B.J. Hiley. *The Undivided Universe: An Ontological Interpretation of Quantum Theory*. Routledge, 1995.

G.E. Bowman. On the classical limit in Bohm's theory. *Foundations of Physics*, pages 605–625, 2005.

H.R. Brown and D. Wallace. Solving the measurement problem: De Broglie-Bohm loses out to Everett. *Foundations of Physics*, 35(4):517–540, 2005.

Detlef Durr and Stefan Teufel. *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer, 2009.

M. Gell-Mann and J.B. Hartle. Classical equations for quantum systems. *Physical Review D*, 47(8):3345, 1993.

R.B. Griffiths. Consistent histories and the interpretation of quantum mechanics. *Journal of Statistical Physics*, 36(1):219–272, 1984.

JJ Halliwell. A review of the decoherent histories approach to quantum mechanics. *Annals of the New York Academy of Sciences*, 755(1):726–740, 1995.

James B Hartle. The quasiclassical realms of this quantum universe. *Foundations of Physics*, 41(6):982–1006, 2011.

P.R. Holland. *The Quantum Theory of Motion: An Account of the de Broglie-Bohm Causal Interpretation of Quantum Mechanics*. Cambridge University Press, 1995.

E. Joos, D. Zeh, C. Kiefer, D. Giulini, J. Kupsch, and I.-O. Stamatescu. *Decoherence and the Appearance of a Classical World in Quantum Theory*. Springer, second edition edition, 2003.

O.J.E. Maroney. The density matrix in the de Broglie-Bohm approach. *Foundations of Physics*, 35(3):493–510, 2005.

M.A. Schlosshauer. *Decoherence and the Quantum-To-Classical Transition*. Springer, 2008.

D. Wallace. *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford University Press, Oxford, 2012.

W.H. Zurek and J.P. Paz. Quantum chaos: a decoherent definition. *Physica D: Nonlinear Phenomena*, 83(1):300–308, 1995.

W.H. Zurek, S. Habib, and J.P. Paz. Coherent states via decoherence. *Physical Review Letters*, 70(9):1187–1190, 1993.

Evidence for causal mechanisms in social science: recommendations from Woodward's manipulability theory of causation

Word count: 4866

Abstract: In a backlash against the prevalence of statistical methods, recently social scientists have focused more on studying causal mechanisms. They increasingly rely on a technique called process-tracing, which involves contrasting the observable implications of several alternative mechanisms. Problematically, process-tracers do not commit to a fundamental notion of causation, and therefore arguably they cannot discern between mere correlation between the links of their purported mechanisms and genuine causation. In this paper, I argue that committing to Woodward's interventionist notion of causation would solve this problem: process-tracers should take into account evidence for possible interventions on the mechanisms they study.

Evidence for causal mechanisms in social science: recommendations from Woodward's manipulability theory of causation

1. Introduction

In a backlash against the pervasiveness of statistical methods (cf. King, Keohane, and Verba 1994), in the last decade certain social scientists have focused on finding the causal mechanisms behind observed correlations (Mahoney 2001, Tilly 2001, Hedström and Ylikoski 2010, Hall 2012). To provide evidence for such mechanisms, researchers increasingly rely on process-tracing, a method which involves contrasting the observable implications of several alternative mechanisms (Bennett and Checkel forthcoming, Brady and Collier 2010, George and Bennett 2005).

The process-tracing methodology literature as of yet does not commit to any particular fundamental notion of causation. Process-tracing reacts to the statistical approach by arguing that finding a correlation between a potential cause and effect variable is not enough evidence for genuine causation. We should also investigate the intervening variables between the putative cause and effect. Process-tracing however does not solve the problem it set out to solve, but rather push the problem one step back. What, after all, is their evidence that the link between these intervening variables are cases of genuine

causation? In this paper, I will show a way out of this problem. James Woodward's manipulability theory of causation tells process-tracers how to find the evidence they need: not only must process-tracers study the intervening variables, but also the *intervention variables* of each link in the causal chain.

This paper is set up as follows. First, I analyse what process-tracing is and what it aims to do. Second, I set out the relevant aspects of Woodward's theory, including my motivation for using his notion of causation rather than another. Third, I evaluate process-tracing in light of Woodward's theory, conclude it indeed lacks evidence for genuine causation, and give recommendations for solving this problem.

2. A philosophical reconstruction of process-tracing

Process-tracing is a mechanism-based method for analysing causal relationships. To be precise, the term refers to two techniques (cf. Bennett and Checkel forthcoming), bottom-up and top-down process-tracing. Bottom-up process-tracing involves surveying a situation of interest with as little preconceptions as possible, in order to then formulate a hypothesis about possible causal connections in that situation. For instance, a researcher may spend time in a post-conflict area in the process of nation-building, interview the population to get data on how secure people feel, and subsequently form a hypothesis about causal links between nation-building efforts and human security. Top-down process-tracing tests type-

level causal hypotheses about the mechanism (the 'process') connecting an independent variable and a dependent variable using data collected in case studies. Bottom-up and top-down process-tracing are occasionally mixed; a researcher may start with a bottom-up study to formulate hypotheses, and continue with a top-down study to see if these hypotheses are corroborated or refuted by the evidence available. In what follows, I will look at the second type of process-tracing, i.e. top-down process-tracing, because I wish to evaluate how process-tracers *justify* causal claims.

First, let us consider top-down process-tracing more formally. The essence of top-down process-tracing is using a case study to contrast rival hypotheses about the causal connection between an independent variable X and a dependent variable Y, that is, hypotheses which suggest rival causal mechanisms that have contradictory observable implications. Let us call the researcher's own hypothesis H_Z. H_Z holds that a causal mechanism Z exists that connects X and Y, i.e. a set of variables Z_i such that $X \rightarrow Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Y$ (where $Z_i \rightarrow Z_j$ means that Z_i causes Z_j). Besides hypothesis H_Z, process-tracers will also investigate any alternative hypotheses H_A, H_B, etc. that are postulated in the literature, that is, they also investigate the observable

implications of chains A or B of intermediate variables $A_1 \ldots A_m$ or $B_1 \ldots B_k$ etc.

Methodologists Andrew Bennett and Jeffrey Checkel argue that the observable implications of mechanisms are "the facts and sequences within a case that should be true if each of the alternative hypothesized explanations of the case is true. Which actors should have known, said, and did what, and when? Who should have interacted with, worried about, or allied with whom?" (Bennett and Checkel forthcoming, 39) According to Sharon Crasnow, process-tracing hypotheses about causal mechanisms are therefore *singular* causal claims, whereas the evidence from case studies consists of *general* causal claims (Crasnow 2012). A philosophical account of process-tracing therefore needs to consider both singular and general causal claims, as well as the link between them. In this paper, I will focus on the general causal claims process-tracing makes. In my conclusion I will, however, indicate what implications Woodward's theory has for the link between singular case study evidence and general hypotheses.

3. Woodward's manipulability theory of causation

Let me now turn to the relevant aspects of James Woodward's manipulability theory of causation, before explicating how we can apply the theory to process-tracing. Woodward

argues that any successful description of a cause-effect relationship must refer to causal factors that can be manipulated to change the phenomenon under study. Specifically, a variable X is a cause of a variable Y if there exists some 'intervention variable' I which we can use to change X, so that X will then in turn change Y without any interference of other variables linked to Y. In other words, using I we can ascertain that X made the change in Y happen.

I have chosen to look at what would happen if the process-tracer committed to Woodward's notion of causation, rather than others, because of two reasons. Firstly, Woodward's theory provides an alternative to the probabilistic notions of causation that are taken for granted in the statistical approaches that process-tracers criticize, such as the one in social science methodology bible King, Keohane, and Verba (1994). Secondly, Woodward's notion is arguably more suited to studying causal mechanisms in social science than the energy-transfer notions of causation developed for causal mechanisms in natural sciences like biology. Thirdly, Woodward's notion has not been widely applied to the social sciences, and therefore this analysis may contribute to the literature in philosophy of causation as well as to philosophy of social science.

Arguably, we could get similar results by accepting Judea Pearl's manipulability framework for causation (Pearl 2000). Though there are formal differences between Pearl and Woodward's notion of an intervention, I believe that my general conclusion in this paper will hold no matter whether the process-tracer commits to Pearl or Woodward's theory.

3.1 Manipulability theory

Let me now outline Woodward's theory. The focal point of Woodward's work is his formal set of necessary and sufficient conditions for X to be a (type-level) cause of Y, which form his manipulability theory:

> A necessary and sufficient condition for X to be a (type-level) **direct cause** of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability of Y when one holds fixed at some value all other variables Z_i in V. (Woodward 2003, 59)

To illustrate the use of the variable set V, consider the following scenario: we are interested in a Scandinavian village, asking whether, for its villagers, eating citrus fruit (X) is a direct cause of an absence of scurvy (Y). To answer that question, we can't just feed the villagers citrus fruit for a month to see what happens to their health. We need to take into account other variables that may influence this (lack of) scurvy. So, we investigate the villagers' diet, and find out that they greatly enjoy eating liver; their liver consumption (Z) is very high. What will happen in our experiments to determine the effect of citrus consumption is the following. If we *ignore* the liver consumption, Z, of the villagers, we will find that no possible intervention on their citrus consumption, X, will change their developing scurvy or not, Y. Simply put, not eating citrus fruit won't mean that the villagers get scurvy. However, *if we keep fixed at 0* the variable Z for these villagers, we will find out that there is an intervention on X, i.e. making the villagers eat citrus fruit, that *will* change Y, i.e. whether they develop scurvy. We find that if $X=0$, i.e. the villagers don't consume the fruit, then $Y=1$, i.e. they develop the deficiency disease. If they *do* consume the fruit, i.e. $X=1$, then they don't develop the disease, i.e. $Y=0$.

The notion of a direct cause alone, however, is too basic for a complete theory of causation. Woodward calls our attention to the possibility of a variable X which influences a variable Y along some route but has no total effect on Y because X's influence is always cancelled out by other factors (Woodward 2003, 50)[1]. In that case, X is not a direct cause of Y, but Woodward nevertheless wants to call X a cause. Therefore he introduces the notion of a contributing cause:

> A necessary and sufficient condition for X to be a (type-level) **contributing cause** of Y with respect to variable set V is that:
>
> (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1,\ldots,Z_n$ such that X is a direct cause of Z_1, which in turn is a direct cause of Z_2

1 This issue is closely related to the notion of 'faithfulness', employed amongst others by Spirtes, Glymour, and Scheines. These 'washing out' cases are cases when faithfulness, defined as being able to read off all causal independence relations off probabilistic (conditional) independence, fails.

, which is a direct cause of … Z_n , which is a direct cause of Y ; and that

(ii) there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value.[2]

If there is only one path P from X to Y , or if the only alternative path from X to Y besides P contains no intermediate variables (i.e., is direct),

2 This second requirement is meant to sort out cases where transitivity of a causal relation fails. To illustrate, such a case, imagine that whilst having breakfast I spill coffee on my navy blazer (C), which causes me to wear a cream blazer instead ($B=c$ rather than $B=n$). Now, it turns out that at my job interview for a fashion editor position that afternoon, wearing a blazer rather than not wearing a blazer (i.e. in this scenario $B=c$ rather than $B=0$¿ causes me to get the job (J) . However, despite requirement (i) being satisfied (after all, there is a directed path $C \rightarrow B \rightarrow J$), we would hardly say that my spilling coffee at breakfast (C) causes me getting the job (J). The causal relation is not transitive. This failure of transitivity is captured by requirement (ii): there is no intervention on my spilling coffee that will change whether I get the job. If I don't spill the coffee, I will wear my navy blazer instead.

> then X is a contributing cause of Y as long as there is some intervention on X that will change the value of Y, for some values of the other variables in V. (Woodward 2003, 59)

As Woodward himself stresses, a direct cause is always a contributing cause, but a contributing cause is not always a direct cause.

3.2 Interventions

The notion of an 'intervention' is a crucial part of Woodward's argument. Note that there is a difference between an intervention variable and a contributing cause variable; whereas a contributing cause variable is part of the situation one is trying to analyse, the intervention variable is the means by which one undertakes this analysis. Before I discuss Woodward's rather technical definition of an intervention variable, I will introduce it with an example.

According to Woodward's theory, introducing a microfinance institution in a country will be an intervention variable I for investigating whether taking out microcredit loans (X) causes a reduction in household poverty (Y) if and only if the following things hold. First, the introduction of the microfinance institution has to increase the probability that a microcredit is taken out. Second, there must be no other source of microcredit loans

besides this microfinance institution (so that when we do not introduce the microfinance institution, no microcredits will be taken out). Third, and this is more difficult to ascertain in practice, the introduction of the microfinance institution should not reduce poverty in a way that is unrelated to microcredits. If it turns out, for instance, that opening a microsavings account also reduces households' poverty, and such accounts are offered by the microfinance institution, the third demand will fail. We would not be able to tell whether the microcredit loan or the microsavings account made the difference. In general, overlooking other ways besides X whereby I may influence Y clouds our judgement about the relation between X and Y. Fourth and last, introducing the microfinance institution must be statistically independent of all variables that reduce poverty by other means than microcredit loans. For instance, if we can only introduce the microfinance institution in regions that have a stable government, this clouds our judgement: the stability of the government could itself cause an eventual reduction in households' poverty. So, we must ascertain that there are no other ways in which I can influence Y; if there were, that would mean that I gives us a misguided picture of the connection between X and Y. (To see the difference between the third and fourth requirement, consider the following. Both the third and the fourth requirement are violated if there is a factor Z causally connected to both I and Y but not to X. Requirement 3 only captures cases in which we have $I \rightarrow Z \rightarrow Y$, whereas for

requirement 4, the relation between I and Z is unknown. It may, for instance, just as well be that $I \leftarrow Z \rightarrow Y$.)

This brings us to the four requirements in Woodward's definition of an intervention for type-level causation:

> " I is an **intervention variable** for X with respect to Y if and only if I meets the following conditions:
>
> (1) I causes X .
> (2) I acts as a switch for all the other variables that cause X . That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I .
> (3) Any directed path from I to Y goes through X . That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y , if any, that are built into the $I-X-Y$ connection itself; that is, except for

(a) any causes of Y that are effects of X (i.e. variables that are causally between X and Y) and
(b) any causes of Y that are between I and X and have no effect on Y independently of X.

(4) I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X." (Woodward 2003, 98)

In short, I is an intervention variable for X with respect to Y when we can use I to check whether X is a (direct or contributing) cause of Y, i.e. when we can use I to change X, where after X will change Y without interference from other variables causally related to Y. Using I, we will be able to ascertain that X made the change in Y happen.

Woodward claims that the intervention does not *actually* need to happen; we may devise a hypothetical experiment. What's more, the intervention does not need to involve human action. A natural process can qualify as an intervention as well. In the microfinance case, it

may well be that there are two regions in the world that are similar in all crucial respects[3] except that one has microfinance institutions whereas the other does not. If we compared the two, taking into account all the requirements above, and found that in the country without microfinance institutions a larger proportion of households was below the poverty threshold than in the country with microfinance institutions, then this would corroborate the claim that there is a causal relation between taking out microcredits and reduction of the proportion of poor households.

Summing up the above, Woodward makes a distinction between contributing causes X, intervention variables I that we use to analyse whether a variable X is in fact a cause, and intervening variables Z that are the means by which a contributing cause X influences its effect Y.

4. Process-tracing evaluated from the perspective of the manipulability theory of causation

3 I concede that this requires one to specify what 'similar' would mean in this context, bringing up such issues as external validity and the reference class problem. I will discuss this in more detail in section 4.1.1.

At first glance, Woodward's notion of a *contributing cause* fits with the hypotheses in a top-down process-tracing study. In what follows I will argue that although the hypothesis H_z has structural similarities with Woodward's notion, nevertheless the proposed methods for testing the hypotheses are quite different. In Woodward's framework, we need to show that all links of the chain connecting X and Y are cases of direct causation, which means we need to show there exists some intervention on X that will change Y. In contrast, all the process-tracing method outlined by methodologists like Bennett and Checkel requires is that we observe the deductive implications of the intervening variables of the mechanism in a case study.

To contrast the two approaches in more detail, consider a simple example. Imagine a social scientist has the type-level causal hypothesis that 'an economic recession, X, is a contributing cause of a drop in non-domestic violent crime, Y, via the intervening

variable of a drop in participation in the night time economy, Z '[4]. In this example, Woodward would urge the social scientist to answer the following questions:

1) Is X a direct cause of Z ? In other words, is there a possible intervention on X that will change Y or the probability of Y when one holds fixed all other variables in V at some value?
2) Is Z a direct cause of Y ? In other words, is there a possible intervention on Y that will change Z or the probability of Z when one holds fixed all other variables in V at some value?

(In practice, as we have seen, this scientist would also investigate the observational implications of alternative mechanisms, e.g. the hypothetical mechanism that X causes Y by means of a rise in the number of people suffering from clinical depression, D .

4 That is, a drop in spending at e.g. pubs and nightclubs. The intuition behind this proposed mechanism is that if someone has less money to spend at pubs and nightclubs they will, all other things being equal, consume less alcohol, and therefore be less likely to partake in violent drunken behaviour, which is one of the main forms of non-domestic violence.

For conciseness' sake, I will not discuss the scientists' study of this alternative mechanism; it will happen analogously to the study of her own proposed mechanism $X \rightarrow Z \rightarrow Y$.)

Using Woodward's definition of an intervention variable (Woodward 2003, 98), we can now adapt our list of required information. To answer question 1, we need to know the following:

1*) There exists a variable I_X which
(1) causes X ;
(2) acts as a switch for X ;
(3) does not directly cause Z and does not cause any causes of Z except those on the path $I_X \rightarrow X \rightarrow Z$;
(4) is statistically independent of any variable A not on the path $I_X \rightarrow X \rightarrow Z$ that causes Z .

and analogously for question 2.

So, concretely, what information does our social scientist need to gather in order to answer to demands 1*) and 2*)? For conciseness' sake, I will focus only on 1*) here, i.e. on finding I_X . The social scientist must find a variable I_X which, firstly, causes the economic recession; secondly, acts as a switch for the economic recession (i.e. makes

whether the recession occurs independent of any other variables); thirdly, does not directly or through a path not on $I_X \rightarrow X \rightarrow Z$ cause the drop in participation in the night time economy; fourthly, is statistically independent of any variable A not on the path $I_X \rightarrow X \rightarrow Z$ that causes the drop in participation in the night time economy. Thus, to find I_X, the social scientist needs to ask herself: could we have prevented the economic recession from happening, in a way that is in no way connected to the night time economy through a different route? And would people participate more in the night time economy if we prevented the recession in this way?

We clearly see the connection between Woodward's interventionist framework and the process-tracers' method break down at this point. A process-tracer interested in the causal connection between X and Y who follows methodologists like George, Bennett, and Checkel is not concerned with finding interventions. Rather, what she does is investigate whether there are observable implications of all three factors (economic recession, drop in participation in the night time economy, drop in total violence) present in some case study. So, she may ask the people living in e.g. the London borough of Hackney whether they go out more or less since the crisis (reasoning that if participation in the night time economy dropped, then these people would confirm they went out less); she may also ask them whether they have experienced violent behaviour (reasoning that if violent crime dropped,

then these people would say that they experienced violent behaviour less often). What she is not required to do, if we take methodological advice from George, Bennett, and Checkel seriously, is come up with an intervention variable. Thus, she will not prove that X, the economic recession, is a contributing cause to Y, the drop in total violence.

On the other hand, whereas a process-tracer can get away with merely noting that the intervening variables in the mechanism are 'instantiated' in some case study, Woodward requires one to supplement this with evidence for *intervention variables* for each link of the mechanism. If we commit to Woodward's framework, this spells trouble for process-tracing methodologists like George, Bennett, and Checkel.

4.1 Woodwardian recommendations for process-tracers

Looking ahead, what advice would the Woodwardian philosopher give to a process-tracer? For both practical and ethical reasons, social scientists may wish to refrain from intervening in social science scenarios themselves. This however does not stand in the way of the Woodwardian process-tracer; as we have seen in section 2, an intervention does not actually need to involve human action. Therefore, one of the ways in which a process-

tracer may find an intervention variable is by looking at two distinct case studies, which are as alike as possible, except in regards to independent variable X .

So, thinking back to our last example, the social scientist should look for a region in which the economic recession did not happen (or perhaps, practically speaking, in which the recession had less of an impact), and another region in which the recession did occur. We must make sure that the reason these two regions differ with regards to the impact of the economic recession is not itself affecting the participation in the night time economy, except through affecting the impact of the economic recession; as an extreme example, if the economic recession had less of an impact on a region because the region was mainly populated by Amish, we would expect the Amish beliefs to be the reason the region's inhabitants do not participate in the night time economy, regardless of the economic recession. Thus, the process-tracer should study at least *two* cases, a 'control case' and an 'experimental case', and justify that these two cases are sufficiently similar.

4.1.1 Sufficiently similar? Causal homogeneity in process-tracing

This brings us to one issue with process-tracing highlighted in the introduction that deserves further study. There is a tension between the methodology of using singular case study evidence on the one hand, and the aim of making general causal claims on the other.

For the social sciences this tension is difficult to resolve: arguably, it requires a causal relevance assumption (cf. Hitchcock 1995) for the set of events generalized over. Christopher Hitchcock argues that singular causal claims (here produced in case study research) can only be used as evidence for a general causal claim if one can demonstrate that the causal relevance of the cause for the effect is the same from individual to individual. Applying this causal relevance criterion to social science, to move from singular case studies to general theories one needs a 'homogeneity assumption': the assumption that in both the target cases and the case under study, ceteris paribus the causal relevance of the cause for the effect is the same.

Consider the following example (following Bakke 2013). Whilst process-tracing political scientist Kristin Bakke looks for evidence for the singular causal claim that the presence of transnational insurgents in Chechnya during the Second Chechen War caused the radicalization of war tactics there, she wants to make the stronger claim that the presence of transnational insurgents in civil conflict more generally causes radicalization. Here, the homogeneity assumption is that in both the Second Chechen War and in all other civil conflicts, all other things being equal the causal relevance of transnational insurgents for the radicalization of war tactics would be the same. That is, the presence of transnational insurgents would raise the probability of radicalization by the same amount in all civil conflicts. This homogeneity assumption is difficult, if not impossible to defend. I do not have time to consider this assumption in more detail here, but note that it deserves further attention.

5. Conclusion

In this paper I have shown that process-tracers generally postulate causal hypotheses which relate a cause X and effect Y by a path Z consisting of intervening variables $Z_1 \ldots Z_n$. They then find a case study in which both C and E are present, and investigate whether $Z_1 \ldots Z_n$ are also present. Woodward defines that X is a *contributing* cause of Y with respect to V if and only if there was a set of intervening variables $Z_1, \ldots, Z_n$ such that X is a direct cause of Z_1, which in turn is a direct cause of Z_2, which is a direct cause of … Z_n, which is a direct cause of Y. As it stands, process-tracing does not establish the complete right hand side of this if and only if statement. Process-tracers show that a set of intervening variables exists, but they do not show that each link of the chain is a relation of direct causation. If they commit to Woodward's notion of causation, process-tracers have to provide evidence that there is a possible intervention to show that the relations they hypothesize are genuinely causal.

References

Bakke, Kristin M. 2013. "Copying and Learning from Outsiders? Assessing Diffusion from Transnational Insurgents in the Chechen Wars." In *Transnational Dynamics of Civil War*, edited by Jeffrey T. Checkel, 31-62. Cambridge: Cambridge University Press.

Bennett, Andrew, and Jeffrey T. Checkel. forthcoming. "Process Tracing: From Philosophical Roots to Best Practices." In *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey T. Checkel. Cambridge: Cambridge University Press.

Brady, Henry, and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd ed. Lanham, MD: Rowan & Littlefield.

Crasnow, Sharon. 2012. "The Role of Case Study Research in Political Science: Evidence for Causal Claims." *Philosophy of Science* no. 79 (5):655-666.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge (MA): MIT Press.

Hall, Peter. 2012. "Symposium: Tracing the Progress of Process Tracing." *European Political Science* no. 11:20-30.

Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* no. 36:49-67.

Hitchcock, Christopher. 1995. "The Mishap at Reichenbach Fall: Singular vs. General Causation." *Philosophical Studies* no. 78:257-291.

King, Gary, Robert O Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, New Jersey: Princeton University Press.

Mahoney, James. 2001. "Beyond Correlational Analysis: Recent Innovations in Theory and Method." *Sociological Forum* no. 16 (3):575-593.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge / New York: Cambridge University Press.

Tilly, Charles. 2001. "Mechanisms in Political Processes." *Annual Review of Political Science* no. 4:21-41.

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Which forms of limitation of the autonomy of science are epistemologically acceptable (and politically desirable)?

4878 words

This paper will investigate whether constraints on possible forms of limitation of the autonomy of science can be derived from epistemological considerations. Proponents of the autonomy of science often link autonomy with virtues such as epistemic fecundity, capacity to generate technological innovations and capacity to produce neutral expertise. I will critically discuss several important epistemological assumptions underlying these links, in particular the "unpredictability argument". This will allow me to spell out conditions to be met by any form of limitation of the autonomy of science to be epistemologically acceptable. These conditions can then be used as a framework to evaluate possible or existing forms of limitations of the autonomy of science. And it will turn out that the option of direct public participation (a lively option in philosophy of science today) might not be the best way to go to democratize the setting of research agenda.

1. Introduction. Pleas for a democratization of the setting of research agenda are often made, and rightly so, on political and moral grounds. In a nutshell, citizens, it is argued, are affected in their daily life by scientific breakthroughs (genetic tests, nanotechnologies, GMO, etc.), and research is (at least partially) funded by their taxes, therefore, in a democratic society, they should have their say in the choices made about research priorities. In itself, this line of argument (which I will endorse here without further arguments) leaves open the issue of which political forms of limitation of the autonomy of science are preferable - a lively option being theses days in philosophy of science some form of direct public participation in the setting of research agenda (e.g. Kitcher 2001, 2011). My aim in this paper is to investigate whether constraints on possible forms of limitation of the autonomy of science can be derived from epistemological considerations.

My starting point will be traditional, utilitarian lines of defense of the autonomy of science. In that perspective, autonomy (in the sense of self-governance) is first considered as a necessary condition for the epistemic and practical successes of science. In other words, when science is left free to define internally its priorities and epistemic aims, it produces more and better knowledge, directly or indirectly useful to society, via in particular technological innovation. Second, autonomy (in the sense of independence and self-regulation) is considered as a necessary condition for the epistemic authority of science. Only when protected from outside influences (commercial, political special interests), so the argument goes, can science deliver the neutral expertise necessary for the proper functioning of a democracy. Note that this link between autonomy and utility for society is at the core of the very influential view of scientific governance defended by Vanevar Bush in his well-know report published in 1945, *Science, The Endless Frontier*. In that document, Bush makes a case for a very broad societal utility of science: "Scientific progress is one essential key to our security as a nation, to our better health, to more jobs, to a higher standard of living, and to

our cultural progress" (1945, 2). And the condition of that progress is, according to Bush, a complete autonomy of science: "scientific progress on a broad front results from the free interplay of free intellects, working on subjects of their own choice, in the manner dictated by their curiosity for exploration of the unknown. Freedom of inquiry must be preserved under any plan for government support of science" (1945, 12).[1]

This kind of utilitarian arguments in favor of scientific freedom associates (more or less implicitly) autonomy with various virtues such as epistemic fecundity, capacity to respond to societal practical needs and neutrality. I will first identify and critically discuss several important epistemological assumptions underlying these links, such as what I will call here the unpredictability argument and the diversity argument. I will distinguish in my discussion autonomy in the sense of self-governance as regards the setting of research agenda and autonomy in the sense of independence and self-regulation, in particular as regards the functioning and composition of scientific communities. This will allow me to spell out conditions to be met by any form of limitation of the autonomy of science to be epistemologically acceptable. These conditions can then be used as a framework to evaluate possible or existing forms of limitations of the autonomy of science. And it will turn out that the option of direct public participation might not be the best way to go to democratize the setting of research agenda.

2. Autonomy and epistemic fecundity. "I didn't start my research thinking that I will increase the storage capacity of hard drives. The final landscape is never visible from the starting point." This statement made by the physicist Albert Fert (2007), winner of the 2007 Noble Prize for his work on the giant magnetoresistance effect, expresses a very common belief, especially among scientists, about the unpredictable nature of the development and

[1] For a more extended discussion of this link between autonomy and societal utility in Bush's report, see for instance Stokes (1997).

results of a research program. Such retrospective observations feed a central argument often invoked in favor of the autonomy, which can be dubbed the *unpredictability argument*. A somewhat lyrical form of this argument was given by Polanyi in his classical essay "The Republic of Science" (1962). Science, says Polanyi (1962, 62), "can advance only by unpredictable steps, pursuing problems of its own, and the practical benefits of these advances will be incidental and hence doubly unpredictable. … Any attempt at guiding research towards a purpose other than its own is an attempt to deflect it from the advancement of science… You can kill or mutilate the advance of science, but you cannot shape it." Therefore scientists must be free "to assess … the depth of a problem and the importance of its prospective solution primarily by the standards of scientific merit accepted by the scientific community." But what exactly is behind this unpredictability argument?

In Polanyi's view, claims about the unpredictable nature of scientific development go hand in hand with a plea for an *internal* definition of research priorities. From this perspective, a problem is deemed important in light of considerations internal to a field of scientific inquiry, such as the potential impacts of its resolution on other epistemic issues central to the field, and not (at least not primarily) in light of external considerations, such as societal utility. The unpredictability argument thus boils down to the claim that because of the unpredictable nature of scientific development, choices of research priorities in a field must only be based on considerations internal to its own dynamic. And if one also buys into the idea that only scientists can master such considerations, then scientists must be left free to define research priorities. But it is easy to see that, as it is, the argument is incomplete. To work in favor of scientific autonomy, the argument must be enriched and reformulated in a comparative form, as follows:

> A field of research is epistemically less productive when its objectives are defined externally than when they are defined internally because, in the latter case, unexpected fundamental discoveries and practical applications are more likely to happen.

This claim immediately raises a simple, empirical question (but which is surprisingly rarely really addressed): is that the case? Does history of science, in particular, show us that a research program whose objectives are defined externally is systematically epistemically less fecund than a research program whose objectives are "disinterested"? After all, examples of finalized research programs having produced along the way unexpected fundamental discoveries are not so rare. For instance the motivations of the Nobel Prize's winners Arno Penzias and Robert Wilson who discovered the 3 Kelvins cosmic microwave background were very practical and had to do at the beginning with improving the quality of transatlantic radio communication. But along the way, this research program led to this very fundamental discovery in cosmology. Industry research on the giant magnetoresistance effect in the 1990s is another telling example of research undertaken under considerable pressure to produce applicable results but which nevertheless produced, along the way, very fundamental knowledge (Wilholt 2006). Of course, an accumulation of examples that could reassure the epistemic pessimistic as regards finalized research will not be enough to invalidate the unpredictability argument, because of its comparative form. Recall that the argument states that a field of research is epistemically *less* productive when its aims are defined externally (i.e. not primarily according to considerations internal to its own dynamic). But the problem is that history of science does not offer any control group. It is just not possible to compare the fecundity of a field when it is left free to define its priority with the fecundity of the same field whose research agenda would be defined externally in order to respond to societal needs. Thus the debate cannot be closed empirically, other considerations are needed.

On the face of it, a promising way could be to draw on what I will call the *diversity argument*. In a nutshell, the argument is the following:

- You cannot predict which line of research (problems + approaches) will turn out to be epistemically fecund or dead ends.
- Maximizing the fecundity of a scientific field thus requires maximizing the diversity of the lines of research (problems + approaches).
- Leaving the scientists free to define their own research agenda is the best way to maximize the diversity of the lines of research.

Clearly, the validity of this argument hinges on the third step, which brings us to considerations partaking of social epistemology. But as far as I know, social epistemology does not provide any good reasons to believe that scientists' freedom of research promotes diversity of lines of research. And given the natural tendency in some fields to monoculture (see for instance the domination of the Big Bang model in cosmology), it seems that we may even reasonably doubt that it is the case[2]. So my first intermediate conclusion is the following: there is no good epistemological reasons to reject any form of externalization of setting of research agenda on the ground that it would diminish the epistemic fecundity of science (*contra* the traditional argument *à la* Polanyi). For all that, claiming that an externalization of the setting of research priorities might be epistemologically acceptable does not mean of course that *any* form of such externalization is epistemologically acceptable. On the contrary, my previous discussion of the link between autonomy and epistemic fecundity establishes that as regards the epistemic productivity of science, what matters is not that

[2] Not everybody would agree though, see for instance Wilholt (2010, 176) for whom the alternative to free choice of projects would be the existence of central authorities who would organized diversity. But it is not clear why an externalization of the setting of research agenda would require that centralized authorities possess "complete and detailed global and local knowledge".

research aims are defined internally, but that the setting of research priorities promotes a diversity of lines of research. Hence the first condition that must be met by any form of limitation of the autonomy of science: ensuring diversity of lines of research.

3. Autonomy and accountability. Let us discuss now a second underlying assumption of the defense of the autonomy of science on utilitarian grounds, that is, the link between autonomy and accountability. The notion of accountability may refer to two distinct types of expectations. Firstly, one can expect from science that it actually delivers the anticipated societal benefits. In other words, to put it trivially, the funding bodies want results for their money. Secondly, expectations can be of a moral nature: scientists can be hold responsible, not only of course for the methods they use, but also for the potential negative impacts their results may have on the citizens' lives[3]. I will focus here on the first kind of accountability, in terms of efficiency to provide the expected societal benefits.

In Bush' views, recall that efficiency is linked to autonomy: autonomy is seen as a necessary condition for science to deliver the expected societal benefits. But is it the case that science is better able to produce what society expects in terms of applicable knowledge and innovations when it is autonomous? Note first that historically, Bush's views started to be called into question on efficiency grounds. Funding bodies, such as the American federal government considered that the return in terms of technological innovations and economic productivity was insufficient. To oversimplify a complex story, American science was considered as too "selfish": too many Noble Prizes and not enough technological innovations (Guston 2000, 138). Challenging the capacity of an autonomous science to actually deliver the expected gains in terms of technological innovations is certainly a good reason to question Bush's views. But I will not discuss further here this delicate and complex issue of the

[3] For a recent discussion of the various dimensions of the notion of scientific responsibility, see Douglas (forthcoming).

efficiency conditions of technological transfer. I would rather draw attention on another essential reason to reconsider the link made between autonomy and efficiency. My proposal is that we should look more closely at the nature of society's expectations and take into account their evolution. For, as noticed by Neal Lane (1997), who was former director of the *National Science Foundation*: "It is not that science did not deliver in so many ways over so many years, but rather that different times require different types of accountability."[4] So what are today the types of societal expectations that science must respond to, and to what extent do they differ from the society's expectations at the heyday of the Bush's model?

I will suggest that the significant feature of the evolution of society's expectations is that they have become more specific, more targeted. First because of the increasing "scientifization" of politics (more and more political decisions call upon scientific expertize on precise issues such as the evolution of the climate or the dangerousness of GMO); second, society's expectations in terms of technological benefits have also become more specific, more targeted. Technological solutions to particular problems are expected (such as how to store photo voltaic energy), and not technological innovation *tout court*, such as the next laser, which would not answer pre-existing needs. So my point is that an autonomous science might well be able to answer global, unspecific expectations (Bush's global expectations - more jobs, better health, technological progress, etc.), but it is very likely less able to meet specific, targeted needs.

My previous analysis of the assumptions underlying the unpredictability argument has shown that a limitation of the autonomy of science (in the form of an externalization of the definition of its research priorities) is epistemologically acceptable as long as it fulfills the condition of diversity of lines of research. The above analysis of the accountability of science in terms of efficiency shows that such a limitation is not only epistemologically acceptable

[4] I borrow this quotation from Guston (2000, 1).

but also necessary. Since societal expectations toward science have become more targeted, an autonomous science whose research priorities are set internally will be less efficient in responding to these expectations.

4. Autonomy and neutrality. Another line of utilitarian defense of the autonomy of science states that only when protected from outside influences (e.g. commercial, political special interests) can science deliver the neutral expertise necessary for the proper functioning of a democracy. Autonomy (here in the sense of independence and self-regulation) is thus considered as a necessary condition for the epistemic authority of science (as long as, of course, this self-regulation obeys proper basic methodological norms). The central issue is then the following: Is a self-governing scientific community more likely to function according to methodological canons that maximize neutrality and impartiality?

Two kinds of considerations may be relevant here: empirical considerations, in that case historical, and considerations provided by social epistemology. Empirical considerations immediately suggest that the condition of self-regulation and independence is far from being enough to guaranty the neutrality of the results produced. Thanks in particular to feminist philosophical and historical studies of science, cases of ideological biases are now well documented in various disciplines. And those cases are not cases *à la* Lyssenko (that is, cases departing from basic methodolgical norms), but cases where a scientific community, largely independent from political power or interest groups, conforming to traditional canons of good science, nevertheless produces non neutral results, which are influenced by dominating ideologies in the broader society. Examples of such "good" but biased science can be found in particular in primatology, archeology, biology, etc.[5] How is that possible? Let me just refer here to Helen Longino's now well-known work, which offers a precise analysis of how

[5] See for instance Keller and Longino (1996).

contextual values may influence the very content of scientific results via the adoption, in the process of empirical justification of hypothesis, of background assumptions. These background assumptions, when they are shared by all members of a scientific community, are invisible and thus avoid the process of mutual criticism at the core a social view of objectivity. Hence the possibility of "good" but biased science. I do not need here to go into more details, for I just want to emphasize one of the main consequences of Longino's analyses in social epistemology, as regards the issue of how a scientific community should be organized in order to reduce the permeability of its results to contextual values. In a nutshell, the (general) idea is that a multiplication of different perspectives on a problem within a scientific community promotes the suppression of biases linked to individual preferences, to the extent that the heterogeneity of viewpoints promotes the identification and the intersubjective critic of the background assumptions involved in the process of empirical justification.

In light of this (rather unproblematic) contribution from social epistemology, our question "Is a self-governing scientific community more likely to function according to methodological canons that maximize neutrality and impartiality?" can be reformulated as follows : does autonomy favor heterogeneity of perspectives within a scientific community? This is admittedly a complex and delicate question. Let me just give here some hints for a negative answer. Scientific communities are not really spontaneously at the cutting edge as regards the social diversity of their composition… So counting on the internal social dynamic of scientific communities to maximize the heterogeneity of perspectives might seem a bit optimistic and even naïve. Therefore, an autonomous scientific community (i.e. not subject to an external control of its composition) might not ensure a very high degree of heterogeneity of perspectives, and thus might not maximize neutrality and impartiality. Some form of external control of the composition of scientific communities, as long as it encourages the

heterogeneity of perspectives, might do better on that terrain. In other words, a limitation of the autonomy of science, in the form of an external control of the diversity of the composition of a scientific community might be necessary to maximize the neutrality and the impartiality of the scientific results produced by these communities, and hence their epistemic authority[6].

Let me just take stoke here. As regards the link between autonomy and epistemic fecundity: the analysis of the validity of the epistemological arguments underlying a defence of freedom of research (in the sense of freedom of choices of research priorities) has established that an externalisation of the setting of research agenda is epistemologically acceptable as long as it fulfils the condition of diversity of research. Analysis of the accountability of science in terms of capacity to respond to societal needs has then established that such an externalization is not only epistemologically acceptable, but also desirable, because of the evolution of the nature of these needs (specific, *targeted* needs are unlikely to be better fulfilled by an autonomous science). As regards now the link between autonomy and neutrality, insights from social epistemology invites to challenge the idea that a self-regulating, self-organized scientific community is better able to produce neutral results: some form of external control of its composition might on the contrary better ensure a heterogeneity of perspectives on a given problem, thus enhancing the neutrality of the results and expertise produced.

5. Evaluative framework. The two aforementioned conditions – condition of diversity of lines of research and condition of heterogeneity of perspectives – provide a framework to evaluate the epistemological acceptability of existing or possible forms of limitation of the autonomy of science. Consider first a form of external control of research priorities already in place and often decried by scientists, to wit, definition of research priorities in light of short-

[6] I borrow from Leuschner (2011) the example of the IPCC as a scientific community whose pluralistically organization is regulated by a political instance (in that case NATO).

term economic interests. In light of the previous analysis, is this form of limitation of the autonomy of science acceptable? The answer is straightforward: such a limitation does not fulfill the first condition of diversity of lines of research and therefore leads to an epistemic impoverishment of science. But note that the problem is not that the objectives assigned to science are defined externally; the problem is rather that these objectives correspond to a very limited subset of the vast collection of objectives assignable to science. In other words, this form of limitation of scientific autonomy should not be rejected on the ground that science should remain "free and disinterested"; it should rather be rejected on the ground that when it comes to the definition of research priorities, considerations of short-term economic profitability should be integrated into a larger collection of considerations, reflecting the diversity of interests, both practical and epistemic, of the *whole* society.

A possible way to realize this integration would be to involve citizens in the choices made on research priorities. This public participation option is indeed widely discussed today and has started to be implemented in scientific institutions, albeit in ways that remain largely anecdotal and purely advisory. In philosophy of science, the ideal of well-ordered science developed by Kitcher (2001, 2011) has become a reference on this matter. In a nutshell, well-ordered science aims at promoting a collective good defined in a non objectivist way, by a process of deliberation involving tutored citizens. This form of direct public participation does indeed offer an alternative to a choice of research priorities in the interests of special groups (such as economic ones): in so far as deliberators are supposed to make evolve their preferences both in light of scientific expertise and in light of others' preferences (hence the notion of "tutored" preferences), the outcomes of the deliberations are supposed to provide an adequate representation of the interests of the *whole* society. For all that, this representativeness does not guaranty that the option of public participation fulfills the two conditions of epistemological acceptability (condition of diversity of lines of research and

condition of heterogeneity of perspectives). As proponents of the autonomy of science would fear, citizens may have rather selective expectations toward science, with a bias toward practical expectations (better cellphones and cures of cancer).

But is this fear grounded? Answering this question would require empirical studies of actual processes of deliberation leading to "tutored" preferences in Kitcher's sense. There exists a relatively large body of literature on consensus conferences and other forms of direct public participation, but these conferences often focus on a particular issue (for instance the societal acceptability of nanotechnology) and not (at least to my knowledge) on the much more general issue of what the research priorities should be at a national or supranational level. In any case, I take the crucial question here to be of a comparative nature: in light of the two conditions of epistemological acceptability, is public participation a form of limitation of the autonomy of science *preferable* to other forms, for instance to some form of control exercised by our elected bodies? My claim is that epistemological considerations do not favor the public participation option over other forms of democratic control. The conditions of diversity of lines of research and heterogeneity of perspectives can also be fulfilled by appropriate forms of control exercised by our elected representatives. One can very well conceived that some appropriate subset of our elected representatives get also "tutored" in Kitcher's sense by scientific experts. Of course, proponents of the public participation option may immediately discard this option on the ground that this would be a far too optimistic view of the capacity of our elected representatives to come up with a large range of scientific priorities, both epistemic and practical, and not just with short-term, economically profitable priorities, under the influence of various powerful lobbyings[7]. But I think this prevention (which may be to a certain extent country-dependant) can be questioned and, in any case,

[7] This seems to be Kitcher's view, for it is striking that governements and elected bodies are completely absent from his picture of a democratized science, or when they are invoked, they are immediately discarded in not very kind terms (2011, 24).

arguments are needed to establish that the public participation option is significantly less prone to biases towards short-term, practical expectations.

6. Concluding remarks. A first conclusion of this paper was that some forms of limitation of the autonomy of science are not only epistemologically acceptable but also desirable. Forms of externalization of the setting of research agenda are epistemically acceptable in so far as they fulfill the condition of diversity of lines of research. And appropriate forms of external control of the diversity of the composition of a scientific community may allow to increase the degree of heterogeneity of perspectives on a given problem, thereby increasing the neutrality of the results produced (see the IPCC example in footnote 7). The next step was then to investigate whether some forms of limitation of the autonomy of science score better than others on these epistemological counts. My claim is that there is no good (epistemological) reason to choose public participation over other forms of democratic control, in particular via our elected representatives. I am very aware that this step has remained very sketchy: much more need to be said to evaluate the comparative merit of the various options of democratic control as regards their capacity to fulfill my two epistemological conditions. In any case, epistemological criteria need to be supplemented by other criteria, such as political representativity, in the sense of "acting for" a larger group (Brown 2004, 86), and integrability within our existing, *representative* systems of democracy. And one can question whether the option of public participation scores well on those two counts. After all, we live in representative democracies where government and elected assemblies are those who are, *in fine*, responsible for the way public money is spent on research. How public participation *à la* Kitcher would articulate with them?

Keeping this issue open, I will just conclude on a general note concerning the type of contributions philosophy of science can bring to the topic of the democratization of science.

In the same way as philosophical reflections on science benefit from taking into account how science actually works, a political philosophy of science should take into account how our democratic systems of decision actually work, as well as the specificities of existing practices in science policies. And given their variety from one country to another, not to the mention supranational levels, this *naturalist* turn will invite a certain degree of localism: rather than trying to come up with a *general* normative proposition on how to democratize science, political philosophy of science should try to elaborate "local" propositions, that is, propositions that take into account the specificities of the relevant institutional and political context and more broadly, the specificities of the relevant "political culture"[8].

References

- Brown, Mark. B. 2004. "The political philosophy of science policy." *Minerva* 42: 77-95.

- Bush, Vanevar. 1945. *Science - The Endless Frontier*. Washington D. C.: National Science Foundation.

- Douglas, Heather. *Forthcoming*. "The Moral Terrain of Science." *Erkenntnis*.

- Fert, Albert. 2007. Interview published in *Le Monde*, October 25, 2007.

- Guston, David.H. 2000. *Between Politics and Science*. Cambridge University Press.

- Jasanoff, Sheila. 2005. *Designs on Nature*. Princeton University Press.

- Keller, Evelyn. F. and Longino, Helen. (eds.). 1996. *Feminism & Science*. Oxford University Press.

- Kitcher, Philip. 2001. *Science, Truth and Democracy*. Oxford University Press.

- Kitcher, Philip. 2011. *Science in a Democratic Society*. Prometheus Books.

- Lane, Niel. 1997. "A devil's paradox: Great science, greater limitations." In AAAS science and technology policy yearbook, 1996/1997. Edited by Albert H. Teich, Stephen D. Nelson,

[8] In Jasanoff (2005) sense of the notion.

and Celia McEnaney, 125-30. Washington, DC: American Association for the Advancement of Science.

- Leuschner, Anna. 2012. "Pluralism and objectivity: exposing and breaking the circle." Studies in History and Philosophy of Science 43(1): 191-198.

- Longino, Helen. 1990. *Science as Social Knowledge*. Princeton University Press.

- Polanyi, Michael. 1962. "The Republic of Science: Its Political and Economic Theory." *Minerva* 1: 54-73.

- Stokes, Donald. 1997. *Pasteur's Quadrant*. Brookings Institute Press.

- Wilholt, Torsten. 2006. "Design Rules: Industrial Research and Epistemic Merit", *Philosophy of Science* 73: 66-89.

- Wilholt, Torsten. 2010. "Scientific freedom: its grounds and their limitations", *Studies in History and Philosophy of Science* 41: 174-181.

I ♥ ♦s

Steven F. Savitt

Department of Philosophy, University of British Columbia

Abstract

Richard Arthur (2006) and I (Savitt 2009) proposed that the present in (time-oriented) Minkowski spacetime should be thought of as a small causal diamond. That is, given two timelike separated events p and q, with p earlier than q, they suggested that the present (relative to those two events) is the set $I^+(p) \cap I^-(q)$. Mauro Dorato (2011) presents three criticisms of this proposal. I rebut all three and then offer two more plausible criticisms of the Arthur/Savitt proposal. I argue that these criticisms also fail.

I ♥ ♦s

1. Causal Diamonds

At the end of the twentieth century, it looked as if one question at the intersection of physics and metaphysics had been settled. What is the present in Minkowski spacetime, M? The upshot of a series of well-known papers beginning in the 1960s seemed to prove that one had a very limited choice. The present, at or for a spacetime point e ∈ M could be either the whole spacetime M or just the point e itself. The choice is no wider if one allows the present to be defined relative to a spacetime point e ∈ M *and* a timelike worldline ɣ containing e.[1]

It might come as a surprise, then, that I (2009) suggested a third structure for the present (relative to e and ɣ) in M.[2] I then called these structures *Alexandroff presents*, but now, to conform to the usage that seems to be standard in physics, I will call them *causal diamonds*. The first order of business must be to define them. Even though the discussion below will mostly concern Minkowski spacetime M, it will be useful to define causal diamonds in a larger class of spacetimes that includes M.

Consider relativistic spacetimes < M,g,↑ > that are strongly causal and possess a temporal orientation (as indicated by the arrow). Choose two points p,q on a timelike

---

[1] The papers from which these ideas emerged were by Howard Stein (1968, 1991) and by Rob Clifton and Mark Hogarth (1995). I will refer to them as *SCH*. These papers were written in response to papers by Cornelis Rietdijk (1966, 1967), Hilary Putnam (1967), and Nicholas Maxwell (1985, 1988). I will discuss the implications of the results in the SCH papers in more detail below.

[2] The same suggestion can be found in Arthur (2006), and a similar idea but to a different purpose in Myrvold (2003, §2). All of us were clearly inspired by the discussion at the end of Stein (1991). One should note also that in the philosophical literature causal diamonds appeared explicitly in Winnie (1997), which in turn was indebted to Robb (1914, 1921, 1936).

worldline ɣ in M with p earlier than q. Then the set $I^+(p) \cap I^-(q)$ is *a causal diamond*.[3] In these spacetimes causal diamonds are guaranteed to exist--for instance, by Theorem 3.27 of Minguzzi and Sánchez (2008). Such spacetimes are free of closed timelike curves, and the topology these sets compose, which is known as the Alexandrov (or Alexandroff) topology, is Hausdorff, giving one what is generally thought to be a physically reasonable spacetime.

Gibbons and Solodukhin (2007a,b) distinguish between small vs. large causal diamonds. Small causal diamonds have a proper time separation between the defining end-points p and q that is small compared to the curvature scale of the ambient spacetime. Larger causal diamonds are those in which the later point q recedes to the future boundary $\mathscr{I}^+$ of an asymptotically de-Sitter spacetime. The cosmologists whose work we will sketch below employ large causal diamonds whereas Arthur and I proposed small causal diamonds (diamonds in which the proper time separation τ between the endpoints p and q is scaled to the human "specious" or psychological present) as (special) relativistic counterparts of the common sense present. But they are all causal diamonds nevertheless.

2. Dorato *contra* Diamonds

Arthur's and my proposal was criticized in Dorato (2011). The aim of this paper is to evaluate these criticisms and then to add a few further thoughts of my own. In the course of this discussion a more detailed understanding of the proposal under fire will emerge.

Dorato crisply sums up his arguments on page 391 of his paper:

---

[3] The set $I^+(p)$ is the set of all points in M that can be reached from p by an everywhere future-directed, continuous timelike curve. The set $I^-(q)$ is the set of all points in M from which a continuous, everywhere future-directed timelike curve can reach q. The set $J^+(p)$ is the set of all points in M that can be reached from p by an everywhere future-directed, continuous timelike or lightlike curve. Similarly for $J^-(q)$. Some physicists think of sets like $J^+(p) \cap J^-(q)$ as the causal diamonds.

(i) [Causal diamonds have] no important applications in physical theories;

(ii) it does not seem a plausible, strong and non-arbitrary explanation of the extendedness of our subjective present, and

(iii) It does not correctly pick out the events we intend to pick out when we use "now" in ordinary language,

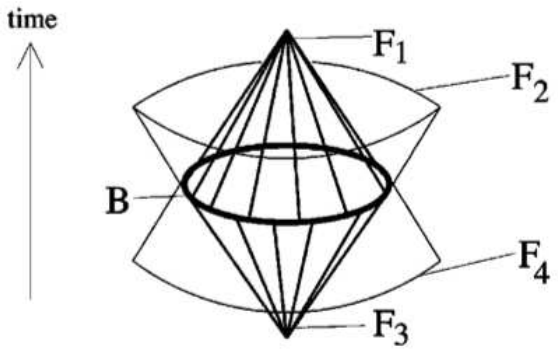
(iv) these seem the only reasons to introduce it.

I conclude that we should drop it.

Let us examine these three criticisms, beginning with the first. My counter-claim is that causal diamonds are well-defined and well-motivated spacetime volumes that have proved, in surprising ways, increasingly handy in recent physics. Let me first advert to authority. Gibbons and Solodukhin (2007a, 2) say that "Causal Diamonds, or Alexandrov open sets, play an increasingly important role in quantum gravity, for example in the approach via casual sets (Sorkin, 2003), in discussions of 'holography', and also of the probability of various observations in eternal inflation models (see Bousso et al., 2007, for a recent example and references to earlier work)." Consider, for instance, holography.

Thomas Banks and William Fischler have been working for a decade or so on a generalization of string theory and quantum field theory they call *Holographic Space-Time* (HST). According to Banks in a recent overview of their work (2013, 2), "The basic geometrical object, for which HST provides a quantum avatar, is a causal diamond… A time-like trajectory can be viewed as a nested sequence of causal diamonds."

To give a simple, related example, let us look at figure 3 of Bousso (2002), a review article on the holographic principle:



The caption of the illustration says this: “The four null hypersurfaces orthogonal to a spherical surface B. The two cones F_1 and F_3 have negative expansion and hence correspond to light sheets. The covariant entropy bound states that the entropy on each light sheet will not exceed the area of B. The other two families of light rays, F_2 and F_4, generate the skirts drawn in thin outline. Their cross-sectional area is increasing, so they are not light sheets. The entropy of the skirts is not related to the area of B.” (Bousso, 2002, 842) The two cones, F_1 and F_3, form a causal diamond. This is only one result of many in the investigation of the holographic principle, but it is one.

The utility of causal diamonds depends on several of their features. First, the volume of a causal diamond is finite, and the area of its boundary is finite. Second, its boundary consists of null or lightlike surfaces. Third, the points in the diamond defined by two points (say p and q) are all those points that can effect some point on a timelike curve extending from p to at q *and* can also be effected by some (other) point on that curve. Bousso imagines an experiment starting at p and ending at q. He claims (Bousso 2000b, especially §2), following Susskind, that physics need take account of only the set of factors that can reciprocally influence the experiment. If so, then physics need consider only events in the causal diamond defined by p and q.

Bousso and Susskind (2011) use causal diamonds for two other purposes. First, they use the boundaries of causal diamonds to define an objective notion of decoherence. When a particle entangled with an apparatus at some event crosses the border of a diamond they define, then (in their view) irreversible decoherence occurs and (in their terms) the event *happens*. Thus they say in §3.3:

> Causal diamonds have definite histories, obtained by tracing over their boundary, which we treat as an observer-independent environment. This gets rid of superpositions of different macroscopic objects, such as bubbles of different vacua, without the need to appeal to actual observers inside the diamond. Each causal diamond history corresponds to a sequence of things that “happen”. And the global picture of the multiverse is just a representation of all the possible diamond histories in a single geometry: the many worlds of causal diamonds!

In addition to providing objective decoherence, Bousso and Susskind, then, use causal diamonds as the many worlds out of which they construct the multiverse in their “multiverse interpretation of quantum mechanics”.

These are some (admittedly, speculative) ways in which causal diamonds have entered into physical theory. I must leave it to the reader to decide whether they are "important". What I would like to emphasize is that causal diamonds are a natural structure to fasten on, since they contain all the spacetime events that can interact causally with events on a timelike worldline γ between the two events, p and q, that define the diamond.

Let me tackle next Dorato's third criticism. Suppose I were to say, on some cold, rainy Vancouver morning, "The sun is surely shining now in Rome." What I would have intended by this (as long as I am not explicitly thinking relativistically) is to pick out events in Rome that are happening at the same time as my utterance and to suppose that those events are part of a sunny day there. To be more pedantic, as far as our common sense, pre-relativistic way of conceiving time goes, my utterance occurs in some observer-independent hyperplane of simultaneous events, and it is meant to signify that the part of the hyperplane that includes Rome contains sunny events.

As I point out (352), but as we all knew already, in the special theory of relativity there is no such distinguished set of simultaneous events. So Dorato is surely right when he says that causal diamonds, if proposed as a scientific successor concept to our common sense concept of the present, do "not correctly pick out the events we intend to pick out when we use 'now' in ordinary language." It is true, however, that *nothing* in M does. Let me just repeat the nice quote from Mermin (2005, xii) that I used to make this point: "That no inherent meaning can be assigned to the simultaneity of distant events is the single most important lesson to be learned from relativity."

So one has to make a choice. Perhaps as far as the special theory goes (and the general theory, insofar as it is locally Minkowskian) there just is just nothing like a (common sense) present to be had in those spacetimes.[4] Alternatively, if one wishes to see what elements of our pre-relativistic concept of time one can find in relativistic spacetimes, one can seek some elements of or structures in Minkowski spacetime (or the more general class of spacetimes stipulated earlier) that *more-or-less* play the role that the common sense present did. If one does make such a proposal, one knows in advance that it will *not* encompass precisely the set of points intended when we use "now" in ordinary language. One looks for a "best fit," with the criteria of fitness rather

[4] I argued this in Savitt (2000).

loosely specified. That is the philosophical task--assuming that there is a philosophical enterprise here at all.

But if that is the game that's afoot, then the suggestion that each event is its own present--no more, no less--certainly has its difficulties. It is not able to assign a truth value to the example above ("The sun is surely shining now in Rome.") spoken by me on the West Coast, although it works well for Dorato in Rome. On the other hand, any reasonably sized causal diamond defined by two events on my world line, one marking the beginning and one the end of my utterance for instance, will include events in Rome and so will afford grounds for assigning a truth value to the example sentence. There will be many, many examples like it. Although the Arthur/Savitt proposal will indeed fail for some other cases (for, say, my musings about what is happening now on Mars), it will do the job in a host of routine situations. I submit that more in the way of correspondence with the common sense present cannot reasonably be asked for in these spacetimes and that therefore Dorato's third criticism is simply beside the point.

Also, if this is the game that's afoot, then Dorato's second criticism above is as wide of the mark as his third. Causal diamonds are not invoked to explain our having experiences of the present that are extended. Rather, our experience of the present as having some duration grounds the requirement (or, more moderately, suggests the possibility) that the relativistic counterpart of the present not be a mere point or an achronal set of points.

In the penultimate paragraph of his paper Dorato says that "violations of achronality are admissible only for the psychological present, but not for the physical present," (393) Viewed one way, this is an eminently sensible view. How could two events that are timelike separated, that are *invariantly* temporally ordered, both be present? But viewed another way, this is the sort of categorical assertion that sometimes comes back to embarrass its author. We live with experienced presents in which a succession of events a second or two long do all seem present, however difficult it may be to articulate this experience coherently. If we are to see what of our commonsense concept of time is afforded to us in relativistic spacetimes, then it is not unreasonable to seek a counterpart of our present that has duration--though, as noted above, it won't be a perfect replica of our commonsense concept. It will be local rather than global, for instance, if Mermin's understanding of Einstein is right.

I conclude that Dorato's three arguments fail. I should stress, however, that even if this claim right, the discussion so far does not show that the Arthur/Savitt proposal is correct. It shows only that certain purported objections are not really impediments to the proposal. There may be other objections to be considered.

3. Region-Relative Becoming

I spoke at the beginning of this paper of theorems that seem to show that the present for a given event in Minkowski spacetime could only be either the event itself or the whole of the spacetime. If that claim is correct, isn't the Arthur/Savitt proposal straightforwardly ruled out?[5] My answer will be: no, I don't think so. How could that be? Well, theorems have conditions, and it may be possible to introduce causal diamond presents by (plausibly) denying one of the conditions of a key theorem. Although the SCH theorems are sufficiently complicated that a full discussion of them is not possible within the available space constraints, it is fortunate that a complete discussion of them is not required. A corollary that contains the material essential for my purpose here was extracted from the SCH results by Craig Callender (2000), and I will restrict my discussion to this corollary.

Let me first just state Callender's "No Go" result. At issue is the definition of a binary relation R, which is intended to represent the relation of "having become". That is, the goal is to define a specific binary relation B such that Bxy holds if and only if y has become with respect to x, where x and y are spacetime points. Stein had proposed (and the proposal seems eminently reasonable) that for such a relation at least all events y in or on the past light cone of an event x should have become as of or for x. Hence condition iii) in Callender's No Go result:

> For any binary relation R on time-oriented Minkowski spacetime, if R is i) implicitly definable from time-oriented metrical relations, ii) transitive, iii) such that, if $y \in J^-(x)$, then Rxy, and iv) satisfies non-uniqueness, then R is the universal relation U. (S592-S593)

[5] Neither Dorato (2011) nor I (2009) discuss this objection.

Condition iv), non-uniqueness, is this:

$$(\text{NU}) \quad (\exists x)(\exists y)(Bxy \,\&\, Byx \,\&\, \sim(x=y))$$

NU, according to Callender (S592), "merely says that at least one event in the universe shares its present with another event's present." Any reasonable representation of becoming should, on this understanding, satisfy condition iii. If two distinct points share a present, as they would in a causal diamond, then it seems that condition iv will be satisfied, and the becoming relation is forced to be the universal relation.[6] This looks to be a disastrous result for any account of the present other than Stein's view that each point event is its own present.

Notice, however, that Callender's gloss on NU contains a metaphysical assumption that, it seems to me, can be reasonably denied. Suppose, for example, that one wished to find an analog for the psychological present in a relativistic spacetime and proposed that some small stretch of a timelike world line γ were the appropriate structure. Then it would turn out that--even given the standard Stein requirement on becoming that we find in condition (iii) of the No Go result and even given the existence of pairs of distinct timelike separated events in that small segment of γ--there would not be two distinct points in that "thick" present that satisfied NU. Having mutually become (which is what NU postulates) is *not* the same relation as "sharing a present."

Similarly a causal diamond will contain (in addition to pairs of timelike separated events) pairs of spacelike separated events x and y such that *neither* Bxy *nor* Byx, but it will not contain events such that *both* Bxy *and* Byx, given the standard Stein condition above. The supposition that the present in a suitable class of relativistic spacetimes can be represented by a causal diamond does not, it seem to me, run afoul of the SCH theorems--unless one requires that events in (or "sharing") a present have become with respect to each other. One need not suppose this, however, if one thinks of the present as a locus of becoming rather than as the "cutting edge" of what has become. Inside that present, events can be partially ordered with respect to becoming in the usual way.

One might wish, however, in addition to the standard Stein definition, to define a notion of becoming relative to that present. More generally, one might wish to define

[6] Of course, I also assuming that the first two conditions are met.

becoming relative to some portion or region of a spacetime, like a causal diamond D. The idea is that what has become relative to D would be all events that have become relative to any event in D, minus D itself. If we call those events B(D), then

$$B(D) \stackrel{def}{=} \{y: (\exists x)\ (x \in D\ \&\ y \notin D\ \&\ y \in I^{-}(x))\}.^{7}$$

One might call this *region-relative becoming.*

If the above defence of small causal diamonds as presents in relativistic spacetimes is successful, it might be argued that I have proved too much. Consider just Minkowski spacetime for the moment. Malament (1977) has shown that, given an inertial world line γ and an event $e \in \gamma$, one can also define the unique hyperplane $\sum$ orthogonal to γ at e. That hyperplane looks very much like the pre-relativistic present, at least as far as the "observer" represented by γ is concerned. B($\sum$) would then be the part of spacetime that has become relative to $\sum$, the past, while the rest of spacetime that is neither $\sum$ nor B($\sum$) is the future relative to $\sum$. Given the naturalness of these ideas, should one not say that $\sum$, rather than D, is the (counterpart of the) present for γ at e in M?

Given the title of this paper, the reader should not be surprised to discover that I think not, but I do not have a knock-down argument for my view. What I can do is offer three considerations that I hope will incline the reader in its favor.

Suppose that two "observers" represented by inertial world lines γ and γ' intersect at some spacetime point e. Both agree as to what has become at e in Stein's sense, $I^{-}(x)$ (or perhaps $J^{-}(x)$). This is a natural and desirable feature of a relation of having become. When it comes to region-relative becoming, however, neither D nor $\sum$ will have this feature. Under reasonable assumptions, however, D will come *very* close.

If the specious presents along γ and γ' centered on an event e are roughly the same temporal length, then their two causal diamond presents (call them D and D') nearly coincide. For each diamond there is a small finite volume of spacetime which will have become relative to one but not the other.[8] The temporal difference of two points in such

[7] Cf. Myrvold (2003, §2).

[8] See the estimate in Savitt (2009, 357-358).

regions (in proper time) will be at most of the order of the proper times of the two specious presents along γ and γ'.

For a pair of hyperplanes $\sum$ and $\sum'$ orthogonal at e to γ and γ' respectively, the case is quite different. There is an infinite volume of spacetime that will have become with respect to each one but not the other, and there is no upper bound on the proper time difference between two points in these regions. D, then, comes *much* closer than $\sum$ to satisfying one desideratum on a notion of the present in the way that it meshes with region-relative becoming.

Secondly, the overlap of D and D' can be used to explain our common sense intuition that at any given time we share a present. The hyperplanes $\sum$ and $\sum'$ have no such large overlap. In the standard presentations of relativity in 1 + 1-dimensional spacetimes, in fact, the only event they have in common is just the point of intersection e.

Thirdly, if one focuses on $\sum$ rather than D in thinking about time in M, then it seems to me that one is willfully ignoring the lesson that one should learn from relativity. Let me quote Mermin again: “That no inherent meaning can be assigned to the simultaneity of distant events is the single most important lesson to be learned from relativity.” There is no reason to choose this one hyperplane as opposed to the infinity of others.

The events in a causal diamond do have an inherent meaning, as thinkers from Alexandrov to Dorato have pointed out. Given an inertial world line containing the events p and q, the causal diamond defined by p and q contains all the events that are “both a possible effect and a possible cause of events on the segment of the worldline [from p to q].” (Dorato 2011, 382) When it comes to understanding time, it might seem odd that the diamonds are local. But our experience is confined to our local region of spacetime, and relativity robs us of justification for extrapolating that experience along an arbitrary hyperplane.

I think these last insights capture at least some of the thought behind my slogan: “Philosophy of time should aim at an integrated picture of the experiencing subject with its felt time in an experienced universe with its spatiotemporal structure.” (351) Dorato protested that the causal diamonds I proposed could not fulfill the expectations raised

by this slogan, and in this he is surely correct. But I did not think that the mere suggestion that one might usefully think of causal diamonds as successor concepts for the present in relativistic spacetimes would complete this program in one go. At best, and if successful, it would be a small first step. It would locate the bits of spatiotemporal structure to be coordinated with the experiencing subjects and with their experiences as one small part of a complex whole that we wish to understand.[9]

[9] I wish to thank Richard Arthur, Adam Brown, Raphael Bousso, Yasunori Nomora and David Rideout for helpful advice. We must all thank Milton Glaser for his gift to New York City. The author of this paper may be contacted at: savitt@mail.ubc.ca.

REFERENCES

Arthur, Richard. (2006). "Minkowski Spacetime and the Dimensions of the Present." in Dennis Dieks, ed. *The Ontology of Spacetime I*. Elsevier.

Banks, Thomas. (2013). "Lectures on Holographic Space-time," arXiv:1311.0755v1 [hep-th].

Bousso, Raphael. (2000a). "The Holographic Principle for General Backgrounds," *Classical and Quantum Gravity* 17:997-1005. [hep-th/9911002].

Bousso, Raphael. (2000b). "Positive Vacuum Energy and the N-Bound," *Journal of High Energy Physics* 0011:038. [hep-th/0010252]

Bousso, Raphael. (2002). "The Holographic Principle," *Reviews of Modern Physics* 74:825-874. [hep-th/0203101].

Bousso, Raphael. and Leonard. Susskind. (2011). "The Multiverse Interpretation of Quantum Mechanics," arXiv:1105,.3796v3. [hep-th].

Callender, C. (2000). "Shedding Light on Time," *Philosophy of Science* 67:S587-S599.

Clifton, Robert. and Mark Hogarth. (1995). "The definability of objective becoming in Minkowski spacetime." *Synthese* 103:355-87.

Dorato, Mauro. (2011). "The Alexandroff Present and Minkowski Spacetime: Why it Cannot do What it has been Asked to Do," in *Explanation, Prediction, and Confirmation: The Philosophy of Science in a European Perspective*, Volume 2. eds. Dennis Dieks, W. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, eds. 379-94. Dordrecht, Heidelberg, London, New York: Springer. DOI 10.1007/978-94-007-1180-8.

Gibbons, G. W. and S. N. Solodukhin. (2007a). "The Geometry of Small Causal Diamonds," *Physics Letters B* 649:317-324. arXiv:0703.098v2. [hep-th].

Gibbons, G. W. and S. N. Solodukhin. (2007b). "The Geometry of Large Causal Diamonds and the No Hair Property of Asymptotically de-Sitter Spacetimes," *Physics Letters B* 652:103-110. arXiv:0706.0603v2. [hep-th]

Jacobsen, Theodore. (1995). "Thermodynamics of Spacetime: The Einstein Equations of State," *Physical Review Letters* 75:1260-63. [gr-qc/9504004v2].

Malament, David. (1977). "Causal Theories of Time and the Conventionality of Simultaneity," *Noûs* 11:293-300.

Maxwell, Nicholas. (1985). "Are probabilism and special relativity incompatible?" *Philosophy of Science* 52:23-43.

Maxwell, Nicholas. (1988). "Discussion: are Probabilism and special relativity incompatible?" *Philosophy of Science*. 55:640-645.

Mermin, N. D. (2005). *It's About Time*. Princeton; Princeton University Press.

Minguzzi, E. and M. Sánchez (2008). "The Causal Hierarchy of Spacetimes," in Alekseevsky, D. and H. Baum, eds. *Recent Developments in Pseudo-Riemannian Geometry*. 299-358. European Mathematical Society Publishing House; Zurich. arXiv:gr-qc/0609119v3.

Myrvold, Wayne. (2003). "Relativistic Quantum Becoming," *British Journal for Philosophy of Science* 54:475-500.

Putnam, Hilary. (1967). "Time and physical geometry." *Journal of Philosophy* 64:240-47.

Rietdijk, C. (1966). "A rigorous proof of determinism derived from the special theory of relativity." *Philosophy of Science* 33:341-44.

Rietdijk, C. (1976). "Discussion: special relativity and determinism." *Philosophy of Science* 43:598-609.

Robb, A. A. (1914). *A Theory of Time and Space*. Cambridge: Cambridge University Press.

Robb, A. A. (1921). *The Absolute Relations of Time and Space*. Cambridge: Cambridge University Press.

Robb, A. A. (1936). *The Geometry of Space and Time*. Cambridge: Cambridge University Press.

Savitt, Steven. (2000). “There’s no time like the present (in Minkowski spacetime)” *Philosophy of Science* 67:S563-S574.

Savitt, Steven. (2009). “The transient *nows*,” in *Quantum reality, relativistic causality, and closing the epistemic circle: Essays in Honour of Abner Shimony*, eds. W. Myrvold and J. Christian. 339-352. Springer.

Sorkin, Raphael. “Causal sets: Discrete gravity” arXiv:gr-qc/0309009.

Stein, Howard. (1968). “On Einstein-Minkowski space-time.” *The Journal of Philosophy* 65:5-23.

Stein, Howard. (1991). “On relativity theory and the openness of the future.” *Philosophy of Science* 58:147-67.

Winnie, J. (1977). “The Causal Theory of Space-Time” in *Foundations of Space-Time Theories, Minnesota Studies in the Philosophy of Science Volume VIII*. Earman, Clark Glymour and John Stachel, eds. 134-205. Minneapolis, Minnesota: University of Minnesota Press.

Killer Collapse

Empirically Probing the Philosophically Unsatisfactory Region of GRW

Charles T. Sebens
University of Michigan, Department of Philosophy

June 23, 2014

Abstract

GRW theory offers precise laws for the collapse of the wave function. These collapses are characterized by two new constants, λ and σ. Recent work has put experimental upper bounds on the collapse rate, λ. Lower bounds on λ have been more controversial since GRW begins to take on a many-worlds character for small values of λ. Here I examine GRW in this odd region of parameter space where collapse events act as natural disasters that destroy branches of the wave function along with their occupants. Our continued survival provides evidence that we don't live in a universe like that. I offer a quantitative analysis of how such evidence can be used to assess versions of GRW with small collapse rates in an effort to move towards more principled and experimentally-informed lower bounds for λ.

1 Introduction

One central point of disagreement in the foundations of quantum mechanics is whether the collapse of the wave function is a genuine physical process. If collapse is to be taken seriously, we should seek to determine physical laws that might govern this process. Ghirardi-Rimini-Weber theory (GRW) offers possible precise laws which guarantee that the wave function collapses during familiar quantum measurements. However, observers and measurements have no special status in the theory, collapses happen all over the place whether or not scientists are watching.

The laws of GRW include two new fundamental constants not present in textbook discussions of quantum mechanics. One parameter, σ, characterizes the precision of the collapse events and the other, λ, the rate at which collapses occur. If these parameters are chosen properly, the theory appears to succeed in generating the correct probabilistic predictions for experiments taken to be within the purview of non-relativistic quantum mechanics. However, as more experiments are conducted we continue to shrink the space of possible values for σ and λ. Potentially, the allowed region could shrink so much it disappears and GRW could be ruled out. Alternatively, new experiments might confirm GRW over its competitors. As of now, there seems to be a fair amount of leeway as to what values we may assign to the parameters (figure 1). Focus on the collapse rate λ. It is fairly well-understood how we can put experimental *upper* bounds

on the collapse rate. If collapse events were too frequent, interference patterns would be destroyed by particles collapsing mid-experiment, isolated systems would heat up, photons would be spontaneously emitted by free particles, and in other varied ways the experimental predictions of the theory would be corrupted (these constraints have been reviewed recently in Adler, 2007; Feldmann & Tumulka, 2012; Bassi *et al.* , 2013).

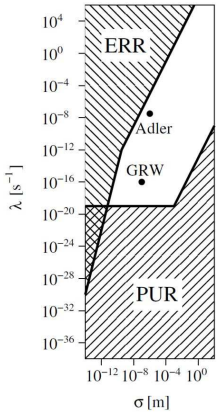


Figure 1: Parameter diagram of GRW theory from Feldmann & Tumulka (2012). ERR is the "empirically refuted region." PUR is the "philosophically unsatisfactory region." The points labeled "GRW" and "Adler" indicate the values suggested in Ghirardi *et al.* (1986) and Adler (2007) respectively. It should be noted that Adler's proposal was made in the context of CSL, not GRW.

In this article, I would like to explore how we might put experimental *lower* bounds on the collapse rate λ. The trend in the literature has been to dismiss low values of λ for non-empirical reasons or for reasons that presuppose the failure of the many-worlds interpretation. When λ is very small GRW becomes an odd theory. Macroscopic objects are not prevented from entering superpositions and the theory takes on a many-worlds character (§3). Such versions of GRW have been rejected as philosophically unsatisfactory. Surely they are. But, there has been disagreement about exactly where the problems arise. Feldmann & Tumulka (2012) give the criterion, "We regard a parameter choice (σ, λ) as philosophically satisfactory if and only if the PO [primitive ontology] agrees on the macroscopic scale with what humans normally think macroscopic reality is like." Bassi *et al.* (2010) impose the requirement that "any superposition reaching the eye must be reduced before it is transformed into a perception in the brain.", building on a suggestion in Aicardi *et al.* (1991). Adler (2007) and Gisin & Percival (1993) argue that the formation of a microscopic latent image in a detector counts as a measurement even before this image is amplified to macroscopic scale. They believe that the collapse rate must be high enough that even these latent images do not enter superpositions.

I will argue that very small values of λ are not just *philosophically* problematic, they are *empirically* unacceptable *even if* the many-worlds interpretation is viable. In doing so, I hope to begin shifting the burden from philosophical considerations to empirical ones and to lay the foundation for a principled and experimentally informed approach to determining lower bounds on λ. Although the paper will focus on GRW throughout, many of the lessons could be applied to *mutatis mutandis* other collapse theories.

2 GRW Theory

In GRW theory, the evolution of the wave function is typically governed by the familiar Schrödinger equation,

$$i\hbar\frac{d}{dt}\left|\Psi(t)\right\rangle = \widehat{H}\left|\Psi(t)\right\rangle \ . \tag{2.1}$$

At some instants, the evolution of the wave function is discontinuous and not in accord with the Schrödinger equation. The wave function collapses. According GRW, collapse is a real physical process governed by well-defined laws and occurring frequently, not just during measurements. Humans and other observers play no spooky role, they are just particularly intelligent and perceptive collections of particles.

When a collapse occurs, a randomly chosen particle has its position become extremely well-localized. Collapses occur randomly at a rate of $N\lambda$ where N is the total number of particles. That is, once a collapse occurs at T_1 the probability that the next collapse, at T_2, will happen within time interval Δt is given by

$$P(T_2 - T_1 < \Delta t) = 1 - e^{-N\lambda\Delta t} \ . \tag{2.2}$$

The collapse rate λ is one of two new constants of the theory, originally suggested to be on the order of 10^{-16}s^{-1} (Ghirardi *et al.* , 1986). The collapse localizes particle I (randomly chosen) around location $\boldsymbol{X}$, where $\boldsymbol{X}$ is chosen randomly with probability density

$$\rho_I(\boldsymbol{x}) = \lim_{t\nearrow T}\left\langle\Psi(t)\right|\Lambda_I(\boldsymbol{x})\left|\Psi(t)\right\rangle \ . \tag{2.3}$$

"$\lim_{t\nearrow T}$" denotes the limit as t approaches the time of collapse, T, from below. $\Lambda_i(\boldsymbol{x})$ is the collapse operator defined by

$$\Lambda_i(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{3/2}}e^{-\frac{(\widehat{\boldsymbol{x}}_i-\boldsymbol{x})^2}{2\sigma^2}} \ , \tag{2.4}$$

where $\widehat{\boldsymbol{x}}_i$ is the position operator for particle i. The wave function after the collapse is given by the pre-collapse wave function multiplied by a tightly peaked three-dimensional Gaussian centered about $\boldsymbol{X}$ and normalized,

$$\lim_{t\searrow T}\left|\Psi(t)\right\rangle = \lim_{t\nearrow T}\frac{\Lambda_I(\boldsymbol{X})^{1/2}\left|\Psi(t)\right\rangle}{\left\langle\Psi(t)\right|\Lambda_I(\boldsymbol{X})\left|\Psi(t)\right\rangle^{1/2}} \ . \tag{2.5}$$

The second new constant in GRW, σ, appears in (2.4) and characterizes the width of the Gaussian that localizes the particle. It was originally proposed to be on the order of 10^{-7}m (Ghirardi *et al.* , 1986). In the remainder of the paper different values of λ will be considered, but σ will be kept fixed at about 10^{-7}m.

In the simplest version of GRW, GRW$\mathbf{0}$, the wave function is all there is and its evolution is determined by the Schrödinger equation (2.1) and the collapse process (2.2, 2.3, 2.5). In the limit where λ is taken to zero, collapse never occurs and GRW$\mathbf{0}$ becomes Everettian quantum mechanics (a.k.a. the many-worlds interpretation or S$\mathbf{0}$). All there is is the wave function and it always evolves in accordance with the Schrödinger equation. Defenders of Everettian quantum mechanics tend to view GRW$\mathbf{0}$ as the right way to

think about GRW theory since they think that our experiences of reality can emerge from patterns in wave functions. For Everettians and others who prefer GRW$\emptyset$ to the alternatives below, this paper can be read as a discussion of GRW$\emptyset$ in the strange regime where it approaches Everettian quantum mechanics.

For some, GRW$\emptyset$ is unsatisfactory (e.g., Allori *et al.* , 2008, §4.3; Maudlin, 2010). According to GRW$\emptyset$ there are no objects in familiar three-dimensional space, just a wave function in an abstract space: a vector in Hilbert space, a complex-valued function on configuration space, or some other exotic beast. In GRWm, the universe contains a wave function which obeys the above dynamics, but that's not all there is, and, in some sense, that's not the important stuff. In particular, it's not the stuff we're made off. In addition to the wave function, there also exists a distribution of matter in three-dimensional space specified by a density,

$$m(\boldsymbol{x}, t) = \langle \Psi(t)| \, \widehat{M}(\boldsymbol{x}) \, |\Psi(t)\rangle \ . \tag{2.6}$$

Here $\widehat{M}(\boldsymbol{x})$ is the mass density operator defined by

$$\widehat{M}(\boldsymbol{x}) = \sum_{i=1}^{N} m_i \, \delta^3(\widehat{\boldsymbol{x}}_i - \boldsymbol{x}) \ . \tag{2.7}$$

In the limit as λ goes to zero, there is no collapse and GRWm becomes Sm, Schrödinger evolution with a mass density (discussed in Allori *et al.* , 2011). Sm is a many-worlds theory much like Everettian quantum mechanics, but where the universe contains a distribution of mass in three-dimensional space in addition to the unitarily evolving wave function. Some think that GRW$\emptyset$ and S$\emptyset$ are unsatisfactory because such laws would not give rise to creatures with conscious experiences like ours, perceiving an apparently three-dimensional world. Readers who think GRW$\emptyset$ is unsatisfactory can understand this paper as a discussion of GRWm in the awkward bit of parameter space where it approaches Sm. In the following sections, I will not differentiate between GRW$\emptyset$ and GRWm. Read GRW in whichever way you think makes it the stronger theory. Read MWI as S$\emptyset$ if you're reading GRW as GRW$\emptyset$, as Sm if you're reading GRW as GRWm.

There is a third version of GRW, GRWf. Here one supplements the wave function with a primitive ontology of flashes. Taking λ to be small in this version of the theory raises entirely different concerns from those faced by GRW$\emptyset$ and GRWm. The problem for GRWf when λ is small is not that human lives are constantly ending, but that such life may be absent altogether. Understanding the empirical adequacy of GRWf in this region of parameter space would require a very different kind of analysis and for that reason GRWf will not be discussed in the remainder of the article. A brief discussion of GRWf in this regime can be found in Feldmann & Tumulka (2012, §4).

3 Branches and Stumps

GRW was originally formulated with the rate of collapse $\lambda \approx 10^{-16}\text{s}^{-1}$. With this rate, when a measurement occurs the wave function just starts to branch into a superposition of outcomes when, with very high probability, the wave function collapses

to a single definite outcome.[1] This is how GRW solves the measurement problem: a definite outcome is guaranteed by the rapid collapse of the wave function and the fact that probabilities for collapsing to different outcomes are given by the Born rule is a non-trivial[2] consequence of the collapse process (2.2, 2.3, 2.5). If the rate of collapse is taken to zero, then collapses never occur and GRW becomes MWI. In MWI, every possible outcome of a quantum measurement actually occurs.

What if λ is chosen so that it is not quite zero, but is very small ($\lambda \ll 10^{-16}\text{s}^{-1}$, keeping $\sigma \approx 10^{-7}\text{m}$[3])? In this regime collapses occur, but only very rarely. When a collapse occurs, the results are catastrophic. After a spin measurement, the laboratory enters into a superposition of a world in which the scientists record an up result and another in which they record down. Later, if any of the particles that compose the scientists or the measurement readout collapse, one of the worlds will be destroyed. Imagine 15 minutes pass between the moment when the measurement occurred and the time when collapse chooses a world to eliminate.[4] In this time, the scientists in both worlds can walk, think, and talk. After collapse, only one world remains. When a collapse like this occurs, all of the inhabitants of the other world are instantaneously and painlessly killed. Or, maybe the collapse doesn't cause the other world to go out of existence, but instead the tail of the Gaussian distorts the world and alters its evolution so that it is inhospitable to human life.[5] In this case, death is quick but perhaps not instantaneous. Either way, in this region of parameter space collapses are not helpful shifts which prevent macroscopic superpositions from forming, they're colossal natural disasters.

The way the universe (a.k.a. multiverse) evolves in each of these three regions of parameter space is depicted in figure 2. With λ at or near zero, worlds branch every time a measurement occurs and each outcome happens on some branch. For standard values of λ, branching is prevented by the collapse of the wave function and each measurement has a definite outcome. For small values of λ branching occurs before collapse is able to prevent it; collapse events occur after branching. Living in such a universe is extremely dangerous as entire worlds are constantly being obliterated. If you are lucky enough to find yourself living a long life, you should be shocked. Repeated improbable occurrences often indicate failure of a theory. This is no exception. The data you receive from your survival provides strong *empirical evidence* against the theory.

[1] There has been some debate over whether the destruction of other branches is successful; see the literature on the problem of tails. Here I assume that the problem can be solved. If it cannot, GRW is not a viable solution to the measurement problem. In particular, I will assume that if collapse chooses one part of the state and massively shrinks the rest, it is not merely improbable to find oneself in a part of the state that was not fortunate enough to be the center of the collapse, it is impossible. There is no life in those other parts after collapse.

[2] For a recent version of the story, see Goldstein *et al.* (2012, §6.5).

[3] This ensures that, in general, a single collapse will be sufficient to destroy branches in which the measurement turned out differently.

[4] This would be typical if we choose λ to be on the order of 10^{-33}s^{-1} and assume that there are about 10^{30} fundamental particles brought into an entangled superposition by the experiment (using (2.2)).

[5] See the brief discussion in Allori *et al.* (2011, §4).

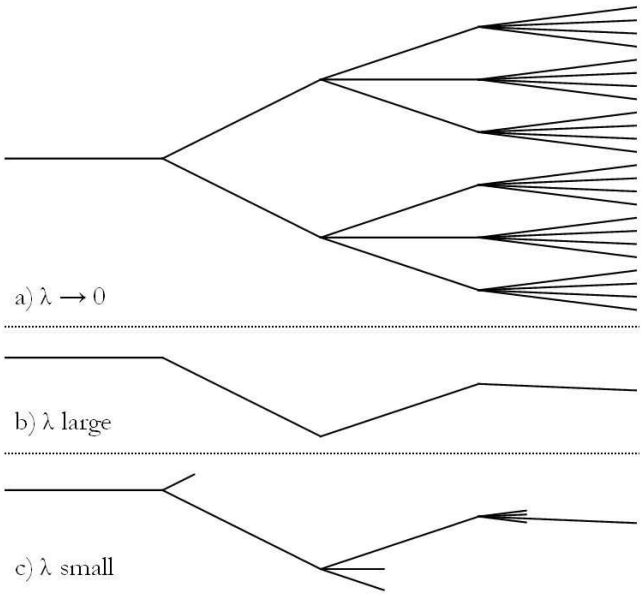


Figure 2: Plot of GRW evolution for a sequence of three measurements for different values of λ.

4 The Rarity of Longevity

To judge the empirical adequacy of a given theory, I will focus on the likelihood of the evidence given the theory, $P(\mathcal{E}|\mathcal{T})$. If, for some evidence $\mathcal{E}$ and theories $\mathcal{T}_1$ and $\mathcal{T}_2$, $P(\mathcal{E}|\mathcal{T}_1) > P(\mathcal{E}|\mathcal{T}_2)$, then the evidence $\mathcal{E}$ confirms $\mathcal{T}_1$ over $\mathcal{T}_2$. If one updates on $\mathcal{E}$ by Bayesian conditionalization, then for any theory $\mathcal{T}$, the credence assigned to $\mathcal{T}$ after gaining the evidence can be expressed in terms of the prior probabilities as $P_{post}(\mathcal{T}) = P(\mathcal{T}|\mathcal{E})$.[6] It follows from the fact that $P(\mathcal{E}|\mathcal{T}_1) > P(\mathcal{E}|\mathcal{T}_2)$ that, if one changes their credences in response to $\mathcal{E}$ by Bayesian updating, the ratio of one's credence in $\mathcal{T}_1$ to their credence in $\mathcal{T}_2$ will rise,

$$\frac{P_{post}(\mathcal{T}_1)}{P_{post}(\mathcal{T}_2)} = \frac{P\left(\mathcal{E}|\mathcal{T}_1\right)}{P\left(\mathcal{E}|\mathcal{T}_2\right)}\frac{P(\mathcal{T}_1)}{P(\mathcal{T}_2)} > \frac{P(\mathcal{T}_1)}{P(\mathcal{T}_2)} \tag{4.1}$$

Theories that are empirically equivalent will assign the evidence equal probability and the data that comes in will not discern between them.

The theories to be compared are: versions of GRW with different parameter values,

[6] Although I expect that this straightforward account of theory confirmation applies to the cases under discussion, one might reasonably be concerned. The situations considered involve *self-locating uncertainty* (see Sebens & Carroll, 2014) and Bayesian conditionalization must be somehow modified to handle such cases (see Arntzenius, 2003). Some modifications will vindicate the use of conditionalization here, others will not. To avoid controversy, I focus primarily on the probability of the evidence given the theory and not the posterior probabilities that result from updating on the evidence.

e.g., $\text{GRW}_{\lambda=10^{-16}\text{s}^{-1}}$; the many-worlds interpretation, MWI; and some unspecified theory which gives the correct Born rule probabilities and guarantees survival, QM.[7] The constraint that QM gives the Born rule probabilities is the constraint that: the probability of seeing the outcome corresponding to eigenvalue O_i of the observable operator $\widehat{O}$ is given by

$$P(O_i|\text{QM}) = |\langle O_i|\Psi\rangle|^2 \ . \tag{4.2}$$

Throughout I'll assume that the agent knows whatever is useful to know about the universal wave function, Ψ, including $|\langle O_i|\Psi\rangle|^2$ for all i. This allows us to focus on the confirmation of alternate dynamical theories without worrying about how agents learn about the universe's wave function.

I will assume that MWI is capable of recovering the Born rule probabilities.[8]

Convenient Conjecture In MWI, after a measurement of the observable $\widehat{O}$ has been made and before outcome is observed, the probability one ought to assign to seeing the outcome corresponding to eigenvalue O_i is given by $P(O_i|\text{MWI}) = |\langle O_i|\Psi\rangle|^2$.

This is a highly controversial assumption, so let me clarify the spirit in which I am introducing it. In order to put empirical lower bounds on λ we need to consider cases where GRW becomes more and more like MWI. If we don't have quantitative predictions from MWI, it will not be possible to quantify the success of GRW in these bits of parameter space. Later I'll discuss how things change when the conjecture is removed (§5).

In the notation used here, $\text{GRW}_{\lambda=0}$ is MWI. So, when a measurement is made, $P(O_i|\text{MWI}) = P(O_i|\text{GRW}_{\lambda=0})$. Thus if we are assuming that the Convenient Conjecture is true and thereby that MWI is empirically adequate, it follows that $\text{GRW}_{\lambda=0}$ is empirically adequate as well.

The question, then, is for what values of λ is GRW approximately empirically equivalent to QM and when do the predictions of GRW and QM diverge? If the predictions diverge significantly, GRW becomes empirically inadequate—the data we actually have fits the predictions of QM. Let's assume for the remainder of this section that the rate of collapse λ is sufficiently small that whenever a measurement occurs we can expect there to be copies of the experimenter who record each outcome. From the Convenient Conjecture and the fact that the dynamics are the same in GRW and MWI before collapse, it is reasonable to suppose that for these small values of λ the probability of seeing each result is given by

$$P\left(O_i|\text{GRW}_\lambda\right) = |\langle O_i|\Psi\rangle|^2 \ . \tag{4.3}$$

But, the observed experimental outcome is *not* the only data one has to update on. The experimenter should also take into account the fact that she has survived for a

[7] What wonderful theory succeeds in recovering the Born rule, as is demanded of the theory I've called "QM"? This will be a matter of disagreement. Let QM stand in for your favorite theory, whichever you think recovers the right probabilities, be it MWI, $\text{GRW}_{\lambda=10^{-16}\text{s}^{-1}}$, Bohmian mechanics, or something else.

[8] For an extended defense of this conjecture, see Wallace (2012). See also Carroll & Sebens (2014); Sebens & Carroll (2014).

time Δt beyond the moment when the measurement was performed. The probability for surviving to Δt can be calculated as

$$\begin{aligned} P\left(\Delta t|\mathrm{GRW}_\lambda\&O_i\right) &= 1 - P\left(\text{fatal collapse by } \Delta t|\mathrm{GRW}_\lambda\&O_i\right) \\ &= 1 - P\left(\text{death}|\text{collapse by } \Delta t\&\mathrm{GRW}_\lambda\&O_i\right) \times P\left(\text{collapse by } \Delta t|\mathrm{GRW}_\lambda\&O_i\right) \ . \end{aligned} \tag{4.4}$$

The probability of a collapse occurring by Δt can be approximated using (2.2) along with the simplifying assumption that there are N_S particles whose collapse would cause a jump to a single outcome: $P\left(\text{collapse by } \Delta t|\mathrm{GRW}_\lambda\&O_i\right) = 1 - e^{-N_S\lambda\Delta t}$.[9] The probability of dying in the event of such a collapse is just the probability that the collapse is centered around some branch other than one's own: $1 - |\left\langle O_i|\Psi\right\rangle|^2$.[10] Inserting these two expressions into (4.4) yields

$$P\left(\Delta t|\mathrm{GRW}_\lambda\&O_i\right) = |\left\langle O_i|\Psi\right\rangle|^2 + e^{-N_S\lambda\Delta t} - |\left\langle O_i|\Psi\right\rangle|^2 e^{-N_S\lambda\Delta t} \ . \tag{4.5}$$

The probability of the total evidence can be assessed by combining (4.3) and (4.5),

$$\begin{aligned} P\left(O_i\&\Delta t|\mathrm{GRW}_\lambda\right) &= P\left(\Delta t|\mathrm{GRW}_\lambda\&O_i\right) \times P\left(O_i|\mathrm{GRW}_\lambda\right) \\ &= \left(|\left\langle O_i|\Psi\right\rangle|^2 + e^{-N_S\lambda\Delta t} - |\left\langle O_i|\Psi\right\rangle|^2 e^{-N_S\lambda\Delta t}\right)|\left\langle O_i|\Psi\right\rangle|^2 \ . \end{aligned} \tag{4.6}$$

We can better understand this formula by considering a simple case. Imagine $\lambda \approx 10^{-33}\mathrm{s}^{-1}$ and $N_S \approx 10^{30}$ so that the experimenter can expect to have approximately 15 minutes between measurement and collapse (as in footnote 4). In this time, she can form expectations about what will happen and look around. Suppose she sees an outcome, O_A, with low Born rule probability, $|\left\langle O_A|\Psi\right\rangle|^2 = \frac{1}{10}$. She should be somewhat surprised and also afraid. Now she knows that she only has a one in ten chance of survival. If she makes it through the day, she should be surprised again. The probability assigned to the total evidence (surviving and seeing that outcome) is $\frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$, which follows from (4.6) with $\Delta t \gg \frac{1}{N_S\lambda}$.[11]

If λ is so small that no collapses are expected to occur within any reasonable length of time Δt and the Convenient Conjecture holds, the predictions of GRW_λ approximately match those of QM. However, as has been noted (Feldmann & Tumulka, 2012, §4), there would be little motivation for such a theory. It would be simpler to

[9] More realistically, N_S would increase as a function of time.

[10] This is an optimistic estimate. In fact there will usually be many worlds corresponding to each outcome and thus even when a collapse is centered on the right outcome O_i, one's world might be destroyed.

[11] What if instead she learns that she's survived before she observes the outcome? Assume for simplicity that there are just two possible outcomes, O_A and O_B. In this case her survival should not be much of a surprise, the probability is 82%. The probability of O_A is 10% and the chance of survival given O_A is 10%. The probability of the other outcome, O_B, is 90% and the chance of survival given O_B is 90%. Thus the total chance of survival is $\frac{1}{10} \times \frac{1}{10} + \frac{9}{10} \times \frac{9}{10} = \frac{82}{100}$. The probability she should assign to O_A given that she survived can be calculated by Bayes' theorem as the probability of survival conditional on O_A, $\frac{1}{10}$, times the probability of O_A, $\frac{1}{10}$, divided by the probability of survival, $\frac{82}{100}$. This yields $\frac{1}{82}$. The probability assigned to her total evidence is the probability of surviving times the probability of seeing O_A upon surviving, $\frac{82}{100} \times \frac{1}{82} = \frac{1}{100}$.

just set λ to zero and remove the collapses all together, yielding MWI. As λ grows it becomes more likely that a collapse will have occurred within Δt and the disagreement between GRW_λ and QM gets worse. QM predicts that you will be alive whereas GRW_λ assigns a certain probability to your death. For fixed λ, the larger Δt is the larger the disagreement between QM and GRW_λ. However, once λ is sufficiently large the crucial assumption that branching precedes collapse becomes invalid. In the next section I'll consider cases in which branching is prevented by collapse.

The fact that one's own continued survival is used as evidence for assessing theories is undeniably odd. Experimenters don't typically keep track of the time elapsed since the experiment was performed. But, epistemologists have contemplated cases much like this where survival *is* relevant data. Consider the following much-discussed example (Leslie, 1989; Swinburne, 1990):

Firing Squad Suppose that a dozen well-trained shooters are ordered to execute you by firing 12 shots each. While blindfolded you hear 144 shots ring out but you survive unscathed.

In such a scenario, your own survival provides evidence that the shooters intentionally let you live over the alternative hypothesis that you got lucky because each of the 144 shots missed its intended target.

The situation here is similar to *Firing Squad*. The hypothesis that the squad intentionally misses is like the hypothesis that QM is true and there are no cataclysmic collapse events. The hypothesis that the shooters were attempting to kill you is like the hypothesis that GRW_λ is true for some troublesome small-but-not-too-small choice of λ where worlds are constantly snuffed out quickly and without warning. However, there is an important difference: In *Firing Squad*, the target will either survive or be killed. In GRW_λ with troublesome λ, there will be many versions of the experimenter that are killed and always at least one that survives. A closer non-quantum analogy is:

Prison Poisoning On New Year's Day you wake up in a nondescript prison cell, #27. A coin was flipped. On New Year's Eve, you were blindfolded and shipped either to *Alcatraz*, if heads, or *Arkham*, if tails. Each prison contains 100 numbered cells and you were randomly assigned to #27.[12] While you slept in your cell the ball dropped and the new year began with a randomly chosen 99 of the 100 cells in Arkham being filled with deadly poison gas. Those in Alcatraz were safe. You knew the plan all along.

In this case, you should initially think it equally likely that you ended up in either prison. After surviving the night you should come to believe that you were probably shipped to Alcatraz since being shipped to Arkham would have likely resulted in your death. It was guaranteed that one of the prisoners in Arkham would survive, but it was not likely to be the one in cell #27. Alcatraz is like MWI and Arkham is like GRW with troublesome λ.[13] The cells represent 100 possible results of a measurement and the gas

[12] For the closest analogy, imagine that each cell of the prison is occupied by a copy of you that resulted from a 1-to-100 fission midday on New Year's Eve.

[13] For an analogue of GRW with a normal collapse rate, consider a prison with a single cell, randomly numbered and free of poison. In this case, the fission in footnote 12 should not be supposed.

plays the role of collapse.[14]

Those who are attracted to the idea of quantum immortality may object to the conclusions reached in this section. Consider a dangerous branching event from the perspective of the many-worlds interpretation (a "quantum suicide" scenario). Suppose you survive on one branch and die immediately, or quickly, on all others. It is tempting to think you should expect survival with certainty. As Lewis (2004) put it, "The experience of being dead should never be expected to any degree at all, because there is no such experience." If death is indeed immediate on all branches but one, the thought has some plausibility. But if there is any delay, it should be rejected. In such a case, there is a short period of time when there are multiple copies of you, each (effectively) causally isolated from the others and able to assign a credence to being the one who will live. Only one will survive. Surely rationality does not compel you to be maximally optimistic in such a scenario.[15] The situation in GRW with a troublesome collapse rate is just like the delayed-death version of the above quantum suicide scenario and, as in that case, survival should not receive probability one. If the collapse rate is raised so that the agent never splits into multiple copies, there is no danger of death and survival can be expected with certainty.

5 Averting Branching

If collapse occurs sufficiently soon after a measurement, branching can be averted. As the other branches of the universe where the outcome was different are just beginning to form, the collapse event occurs, ensuring that the macroscopic readout gives a definite result and the experimenter sees a single outcome. The simplest way to implement this feature of the theory is by imposing a cutoff characterizing the amount of time that passes before branching occurs if there is no collapse. If a collapse happens within τ, branching is averted and a single outcome occurs. If collapse does not occur until after τ, then there is a branching of worlds before the collapse, as in the previous section. Let $C_{<\tau}$ indicate that collapse occurs before the cutoff, $C_{>\tau}$ indicate after. Including both of these possibilities, the probability of the data given the theory can be expressed as

$$P\left(O_i\&\Delta t|\mathrm{GRW}_\lambda\right) = \overbrace{P\left(O_i\&\Delta t|\mathrm{GRW}_\lambda\&C_{>\tau}\right)}^{①} \times \overbrace{P\left(C_{>\tau}|\mathrm{GRW}_\lambda\right)}^{②} + \underbrace{P\left(O_i\&\Delta t|\mathrm{GRW}_\lambda\&C_{<\tau}\right)}_{③} \times \underbrace{P\left(C_{<\tau}|\mathrm{GRW}_\lambda\right)}_{④} . \tag{5.1}$$

The first piece, ①, is just as in (4.6) where it was assumed that branching preceded collapse. The fourth piece, ④, is the probability that a collapse happens by τ. This

[14] Cases like *Prison Poisoning* and *Firing Squad* have a curious feature: one hypothesis cannot be confirmed by the subject in the scenario. If the poison acts instantly, no course of experience would support the Arkham hypothesis over Alcatraz. Similarly, if collapse kills instantly there are no experiences one could have that would provide evidence for GRW with troublesome λ over QM (*if* the Convenient Conjecture holds).

[15] The situation here is like that of the prisoner in Arkham if the period between the splitting event (see footnote 12) and the deaths were made much shorter.

follows directly from (2.2), ④ $= 1 - e^{-N_S \lambda \tau}$. The second piece is simply the probability that a collapse does *not* occur, ② $= 1 -$ ④. The third piece, ③, is the probability that a given outcome resulted from the GRW collapse process in a case where branching does not occur. Here we have GRW working as intended and the probability should be in approximate agreement with the Born rule provided λ is not so large as to push us into the empirically refuted region of parameter space (figure 1), ③ $\approx |\langle O_i|\Psi\rangle|^2$. Inserting these expressions in (5.1) and rearranging gives,

$$P\left(O_i \& \Delta t | \mathrm{GRW}_\lambda\right) = |\langle O_i|\Psi\rangle|^2 - \Big(1 - |\langle O_i|\Psi\rangle|^2\Big)\Big(1 - e^{-N_S \lambda \Delta t}\Big)|\langle O_i|\Psi\rangle|^2 e^{-N_S \lambda \tau}\ , \tag{5.2}$$

which limits to the Born rule probabilities as λ goes to zero or infinity.[16] (5.2) is not valid if λ is large enough that the probabilities in ③ deviate significantly from those given by the Born rule. It cannot be extended in a simple and general manner as the way in which ③ deviates from $|\langle O_i|\Psi\rangle|^2$ will be depend on the particular experiment under consideration.

To recap: If λ is so extremely small that you should not expect (relevant) collapses to have occurred in your lifetime (figure 2.a), then GRW_λ is empirically adequate if the Convenient Conjecture holds. If λ is large enough that collapses must be considered but small enough that branching typically precedes collapse (figure 2.c), then early death is the norm and one's continued survival provides strong evidence against the theory. If λ is increased to around the initially proposed value of $10^{-16}\mathrm{s}^{-1}$ (figure 2.b), the theory may again be empirically adequate as branching is prevented by collapse and the collapse process ensures that the probabilities of various outcomes are given by the Born rule. If λ is increased even further, so that $\lambda > 10^{-8}$, the theory is again empirically inadequate as collapses occur too frequently. Superpositions are destroyed mid-experiment and other maladies ensue (see Feldmann & Tumulka, 2012; Bassi *et al.* , 2013).

What happens if the Convenient Conjecture is false and MWI gives different probabilities from QM? Then, $\mathrm{GRW}_{\lambda=0}$ is empirically inadequate as $\mathrm{GRW}_{\lambda=0}$ is MWI. This failure also rules out GRW_λ for very small λ where collapses can be neglected. For larger values of λ where collapse is rare but non-negligible, there are now two ways in which the theory fails: the probabilities of the various outcomes are incorrect and there is, in general, some probability that one would not have survived to Δt. For still larger values of λ that successfully avert branching, the theory again has a chance of being empirically adequate since the probabilities of outcomes are now determined by the collapse process and the MWI probabilities are irrelevant.

[16] In this simplified story, the probability of surviving to Δt and seeing a certain outcome O_i depends dramatically and discontinuously on whether collapse happens before or after branching. The expressions for ① and ③ are quite different. A more careful analysis would ideally give a smooth transition, but this would require wading into the murky territory of collapses that occur *during* branching (as branching is gradual not instantaneous) and settling questions of personal identity there (in particular, when exactly personal fission occurs and whether it can, in any relevant sense, partially occur). It might be seen either as intriguing or disconcerting that we must answer questions of personal identity to put precise lower bounds on λ.

6 The Race: Decoherence vs. Collapse

For GRW to be tenable, there must be values of λ for which the theory is empirically adequate. On the one hand, λ must be large enough that collapse practically never occurs after the experimenter has branched into multiple copies. Otherwise, one's continued survival empirically refutes GRW_λ, (5.2). On the other hand, λ must be small enough that collapses do not spoil the results of experiments that have been performed. That is, λ must lie below the experimentally refuted region of figure 1. But, are there any values in this range? To answer this, we need to determine whether decoherence-induced branching tends to occur before or after collapse.

We know that for values of λ near the originally suggested value, 10^{-16}s^{-1}, the experiment readout and the experimenter are in a well-defined state corresponding to a single outcome very soon after the measurement occurs. But, what is not clear is which of two possibilities occurred immediately after the measurement (figure 3): (a) the world briefly branched and then a collapse event destroyed some of the copies of the experimenter, or (b) there was never a branching event because collapse prevented the microscopic superposition from causing the experimenter to enter into a superposition.

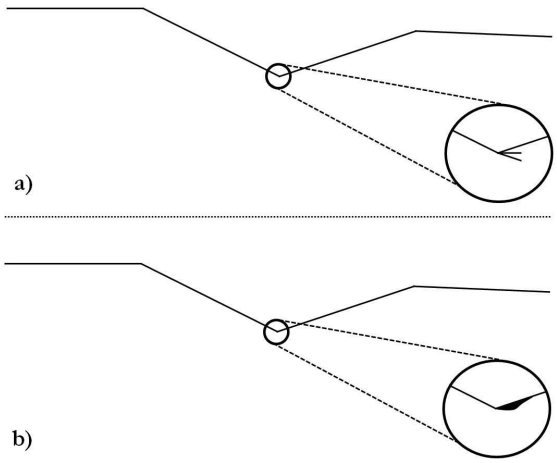


Figure 3: Two potential close-ups of figure 2.b.

A proper analysis is warranted, but beyond the scope of this paper. Here is a *very* rough calculation of how quickly collapse would have to occur to prevent decoherence-induced branching: Decoherence is fast. A slow estimate might be 10^{-23}s for 1 gram of matter at room temperature in a superposition of two locations separated by one centimeter (Zurek, 2003). To ensure a 95% probability of collapse by 10^{-23}s, λ would have to be at least $3\ \text{s}^{-1}$ (from (2.2), assuming the number of particles is on the scale of moles, $N = 10^{23}$). But, experiments restrict λ to being at most 10^{-8}s^{-1} (figure 1). This calculation suggests trouble. There may not be a safe region of parameter space.

Let me highlight two of the most pernicious simplifications in this rough calculation: First, it is assumed that the bit of matter starts in a superposition. In actuality, it

would take time for the matter to enter a superposition and a collapse event could occur in this interval, preventing the macroscopic superposition from forming. Second, when decoherence occurs in this scenario one may doubt whether there is a branching of worlds and in particular whether the experimenter branches. In GRW**0**, it's tempting to say there that the experimenter has branched as there are now two well-separated parts of the wave function that will never again interact (non-negligibly), even if no future collapses occur. In GRWm, it is easier to resist this conclusion as the mass-density of the experimenter may be unaffected by the decoherence of this macroscopic object.

I'll close by summarizing the key lessons of the analysis. First, to determine precise experimental bounds on the parameters λ and σ in GRW, we must determine the probabilities assigned to different outcomes in MWI (§4). This provides additional motivation for that ongoing research program. Second, even if the Convenient Conjecture holds and MWI is empirically adequate, some of the philosophically unsatisfactory region of parameter space is also empirically refuted (§3, 4, & 5). Surprisingly, it is not refuted by the outcomes we observe, but by the fact that we live long enough to observe so many of them. Third, it is not clear how to draw a principled border for the philosophically unsatisfactory region if our dissatisfaction is purely "philosophical" (§1). But, with the realization that small values of the collapse rate λ are empirically refuted, we now have a method to begin drawing principled lower bounds on λ: determine whether the experimenter branches before or after collapse (§5 & 6). Simple calculations suggest that the lower bound generated from empirical considerations will be stronger than the bound generated from a distaste for long lasting macroscopic superpositions, perhaps strong enough to rule out GRW entirely (§6). This merits further study.

7 Acknowledgments

Thanks to David Baker, Gordon Belot, J. Dmitri Gallow, Jeremy Lent, David Manley, Laura Ruetsche, and Roderich Tumulka. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 0718128.

References

Adler, Stephen L. 2007. Lower and upper bounds on CSL parameters from latent image formation and IGM heating. *Journal of Physics A: Mathematical and Theoretical*, **40**(12), 2935.

Aicardi, Franca, Borsellino, Antonio, Ghirardi, Gian Carlo, & Grassi, Renata. 1991. Dynamical models for state-vector reduction: do they ensure that measurements have outcomes? *Foundations of Physics Letters*, **4**(2), 109–128.

Allori, Valia, Goldstein, Sheldon, Tumulka, Roderich, & Zanghì, Nino. 2008. On the Common Structure of Bohmian Mechanics and the Ghirardi–Rimini–Weber Theory. *The British Journal for the Philosophy of Science*, **59**(3), 353–389.

Allori, Valia, Goldstein, Sheldon, Tumulka, Roderich, & Zanghì, Nino. 2011. Many Worlds and Schrödinger's First Quantum Theory. *The British Journal for the Philosophy of Science*, **62**(1), 1–27.

Arntzenius, Frank. 2003. Some problems for conditionalization and reflection. *The Journal of Philosophy*, **100**(7), 356–370.

Bassi, Angelo, Deckert, Dirk-André, & Ferialdi, Luca. 2010. Breaking quantum linearity: Constraints from human perception and cosmological implications. *EPL (Europhysics Letters)*, **92**(5), 50006.

Bassi, Angelo, Lochan, Kinjalk, Satin, Seema, Singh, Tejinder P, & Ulbricht, Hendrik. 2013. Models of wave-function collapse, underlying theories, and experimental tests. *Reviews of Modern Physics*, **85**(2), 471.

Carroll, Sean M., & Sebens, Charles T. 2014. Many Worlds, the Born Rule, and Self-locating Uncertainty. *In:* Struppa, Daniele, & Tollaksen, Jeff (eds), *Quantum Theory: A Two-Time Success Story: Yakir Aharonov Festschrift*. Springer.

Feldmann, William, & Tumulka, Roderich. 2012. Parameter diagrams of the GRW and CSL theories of wavefunction collapse. *Journal of Physics A: Mathematical and Theoretical*, **45**(6), 065304.

Ghirardi, GianCarlo, Rimini, Alberto, & Weber, Tullio. 1986. Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, **34**(2), 470.

Gisin, Nicolas, & Percival, Ian C. 1993. The quantum state diffusion picture of physical processes. *Journal of Physics A: Mathematical and General*, **26**(9), 2245.

Goldstein, Sheldon, Tumulka, Roderich, & Zanghì, Nino. 2012. The Quantum Formalism and the GRW Formalism. *Journal of Statistical Physics*, **149**(1), 142–201.

Leslie, John. 1989. *Universes*.

Lewis, David. 2004. How Many Lives has Schrödinger's Cat? *Australasian Journal of Philosophy*, **82**(1), 3–22.

Maudlin, Tim. 2010. Can the World be Only Wavefunction? *Pages 121–143 of:* Saunders, Simon, Barrett, Jonathan, Kent, Adrian, & Wallace, David (eds), *Many Worlds?: Everett, Quantum Theory, & Reality*. Oxford University Press.

Sebens, Charles T., & Carroll, Sean M. 2014. Self-Locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics. arXiv:1405.7577 [quant-ph].

Swinburne, Richard. 1990. Argument from the Fine-Tuning of the Universe. *Physical cosmology and philosophy*, 154–173.

Wallace, David. 2012. *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford University Press.

Zurek, Wojciech H. 2003. Decoherence and the transition from quantum to classical–REVISITED. quant-ph/0306072.

You Can't Go Home Again – or Can You? 'Replication' Indeterminacy and 'Location' Incommensurability in Three Biological Re-Surveys

Ayelet Shavit, Tel Hai College, Israel

Abstract

Reproducing empirical results and repeating experimental processes is fundamental to science, but is of grave concern to scientists. Revisiting the same location is necessary for tracking biological processes, yet I argue that 'location' and 'replication' contain a basic ambiguity. The analysis of the practical meanings of 'replication' and 'location' will strip of incommensurability from its common conflation with empirical equivalence, underdetermination and indeterminacy of reference. In particular, I argue that three biodiversity re-surveys, conducted by the research institutions of Harvard, Berkeley, and Hamaarag, all reveal *incommensurability* without *indeterminacy* in the smallest spatial scale, and i*ndeterminacy* without *incommensurability* in higher scales.

Key words: replication, location, biodiversity, incommensurability, indeterminacy and empirical equivalence.

Acknowledgment: I deeply thank the people who's ideas and personal example inspired this work – Yemima Ben Menahem, Elihu Gerson and James Griesemer – and the organization that supported it: the Israeli Science Foundation (ISF grant no. 960/12).

1) Replication

Replication - "the set of technologies which transforms what counts as belief into what counts as knowledge" (Shapin and Schaffer 1985, 225), is fundamental to science. Repeatability of a scientific experiment and reproducibility of its results is a common scientific practice ever since Boyle (1660/1999)[1] and Redi (1668/1909), and is widely accepted that one cannot fully explain a biological process nor empirically confirm a generalization without it (Shavit and Griesemer 2009; Shavit 2013).

Philosophers of science were traditionally more skeptical of the possibility and relevance of replication. Problems concerning replication were initially presented as epistemic absurdities, from Wittgenstein's 1953 rule-following paradox: "No course of action could be determined by a rule because every course of action could be made out to accord with the rule" (Ibid. I: 201); to Popper's note of the relativity of similarly, "But if repetition is thus based upon similarity...[it] means that anything can be made to a repetition of anything, as long as we adopt the appropriate point of view" (Popper 1959, 422), to Collins's 1985/1992 experimental regress: "The problem is that, since experimentation is a skillful practice, it can never be clear whether a second experiment was conducted sufficiently enough to be considered as check on the results of a first. Some further test is needed to test the quality of the experiment - and so forth (Ibid. 2). Hacking (1983) concludes that the concern with replication is a philosophical pseudo-problem

[1] The challenge of replication may date back to Heraclitus's metaphor of stepping into the same river twice (Hearclitus DK22B91, DK22B12 translation: Robinson 1987), yet such straightforward comparisons are clearly problematic (Hadot, 1995/2002.)

"…because, roughly speaking, no one ever repeats an experiment (Ibid. 231)".

Given this long tradition of skepticism, it is apparently surprising to learn about the scientists' widespread and genuine concern, or what *Nature* editors referred to as "the plague of non-reproducibility in science" (Hayden 2013): the fact, that widely-published research in many scientific fields is never replicated, and may not even be replicable nor become generalizable (in "Nature" see: Bissel 2013; Baker 2012; Gun 2014; Russell 2013; Sanderson 2012). The bulk of attention is focused on biomedical research, but owing to the overwhelming variability in scope, scale, data structure, and semantics for studying the dynamics of our environments, the problem of reliable replication is clearly applicable to ecological and biodiversity research (Michener and Jones 2011), as well as agriculture, molecular biology, bioinformatics and other biological disciplines (Shavit and Ellison (eds.) 2014 accepted for publication). Furthermore, in biological research, the spatial and temporal contexts – the *location* of a genome, cell, organism, population, habitat or ecosystem – as well as the researcher's questions, methods, and available means of funding are constantly changing. Since biological research is contingent on the historical and social context in which it is being conducted, biologists are confronted with this key *challenge*: how do we both conceptualize and implement (operationalize) replication?

The term 'replication' refers to wide-ranging practices: a) *repeating* the same exploration process (sampling, experimentation, and so on), and obtaining comparable results; b) *reproducing* the same result from the same analysis on the resultant data without a new exploration process (Cassey and Blackburn 2006); and c) *retrieving* the individual entity, a physical item (specimen, photo, blood tissue etc.) or a data record (stored in a field-journal,

excel spreadsheet, SQL database and so forth) for aggregating, comparing, and interoperating the data.[2] In the section bellow, I will make the case for an unavoidable ambiguity that will prevent the *repeatability* of a biodiversity survey at the smallest spatial scale. The third section explains why that much ambiguity does not threaten the *reproducibility* of biodiversity data in higher spatial scales, as long as scientists are mindful of the serious problem of repeatability and therefore record the wider context and the history of their work, as they make this context easily and automatically *retrievable*.

2) Location

Ever since biodiversity (or "scientific natural history") became engulfed in a range of scientific disciplines (Kohler 2006, 2012; Strasser 2008), revisiting the same location was necessary for the scientific study of ecological systems sensu lato (Latour 1999 Ch. 2). Any explanation in ecology, biogeography, or biodiversity requires at the minimum, an identified location, a description of the distribution patterns of a population or species, and a comparison of location patterns for one or more spatial scales. Variables that correspond to changes in pattern, such as the location's average temperature, may therefore point to the process or processes that caused such a change in the distribution of organisms and groups of organisms (populations, species etc.). In response to a global climatic change (Lloyd 2010) and a global crisis of species' extinction (Willson 1992, Ch.12), the biological

[2] "Interoperability is the ability of two or more systems or components to exchange information and to use the exchanged information." IEEE (Institute of Electrical and Electronics Engineers) 1990, 42.

community was intensely engaged in meticulously tracing a species' location back to its geographical position (Tingley 2009), as they made sure that their information will be kept available and interoperable for others or for individual use in the future (Bowker 2005; Ellison et. al 2006).

The inherent vagueness of 'location' is discussed in depth, although in very different contexts, in the philosophy of quantum physics (Barad 2007), in Science and Technology Study of maps (Black 1997; Gugerli 1998) and in eco-feminist studies of the politics of inscribing places (Shiva 2000; Code 2006). However, a study of the various non-metaphorical meanings of 'location' on multiple scales is relatively new to the philosophy of biology (Shavit and Griesemer 2009, 2011ab).

In order to clarify the concepts of 'location' and 'replication', in addition to literary analysis, three case studies involved a philosopher, who for at least three years participated in fieldwork, lab meetings, and workshops. The case studies involved the research institutions of Harvard Forest, Harvard University, the Museum of Vertebrate Zoology (MVZ), Berkeley and Hamaarag, and Israel's Academy of Sciences – which have conducted rigorous repeated surveys to designated locations across New England, California, and Israel respectably. The concepts, working protocols, and conclusions of these case studies set national and international standards in biodiversity research (Shavit 2013); hence the analysis of their use of 'location' and 'repeated sample' and their debates is expected to be highly relevant for science and the philosophy of science alike.

The problem is that two concepts of space – exogenous and interactionist – are each committed to different epistemic values and standards of replication, and are both

necessary for a rigorous repetition of a survey to the same location. An "exogenous" concept of space assumes that organisms' impact their environment - through their physiology, metabolism, behavior, or sheer existence - can be safely ignored for successfully predicting their distribution (Hutchinson 1978, 159-60; Guisan and Thuiller 2005). An alternative "interactionist" concept of space raises the assumption that organisms and their environments are mutually co-determined (Levins and Lewontin 1985, 1-2).[3]

Adopting a specific concept of space signifies a commitment; an actual expenditure of resources (Gerson 1998) to specific constitutive and contextual values, cognitive and social constrains (Longino 2004), and to generatively entrenched (Wimsatt 2007, Ch. 7) work procedures for coordinating the scientific work (Gerson 2007). An exogenous concept of space is committed to revealing general distribution patterns, hence values *representative* data. On the other hand, an interactionist concept of space presumes that one cannot typically ignore organism-environmental interactions, as sometimes they make a relevant casual difference in a species' location, hence values *comprehensive* data on that particular location and species.

[3] Other philosophical traditions that explore the codetermination of organism and environment include the developmental systems theory (Oyama 1985/2000 ;Griffiths and Gray 1994), the scaffolding perspective (Griesemer 2014) and Sterelny's (2001) environmental engineering approach. The biological literature an interactionist concept of space reveals in niche construction (Odling-Smee et al. 2003), foundational-species (Ellison et. al 2010) and eco-engineer (Jones et al. 2007) models.

Given the goal of representativeness, an exogenous partition of space strives to locate a measuring device (e.g. climatic chamber, trap, camera, etc.) on a preselected random point that defines the longitude, latitude and angle of a regular shape (e.g. rectangle, hexagon, transact line, etc.) and deliberately attempts to ignore any hypothetical prior knowledge on the historical and biological context of the species, location and the studied field. On the other hand, an interactionist partition of space seeks to set that exact device in an irregular polygon to form according to a preselected, non-random environmental stratification (e.g. microhabitats, patch-type, participation gradient etc.) hypothesized to be relevant for understanding the dynamics of a particular biological system. The scientists in all three cases, who jointly wrote the research grant and/or agreed on the sampling method only days before venturing outdoors, were surprised to learn that these practices became *mutually exclusive*. It raises the question: what to do first? An exogenous protocol for identifying a location requires to randomly[4] preselect longitude, latitude, depth/elevation and an angle for an individual's measuring device, and only after its point-location was established and entered for recording its microhabitat surroundings. An interactionist protocol requires the opposite: to firstly identify outdoors the location of a preselected microhabitat suspected of being casually relevant to the species and/or the environment, then set up the measuring device within/outside of the micro-habitat, and only then record its lat./long. coordinates.

Since both concepts of space are necessary for rigorous biodiversity surveys, each

[4] Most sampling is haphazard, uniform or hierarchical rather than purely randomized (Shrader-Frechette and McCoy 1993).

concept however binds the researcher to different work practices for maintaining its standard, and since one cannot utilize both procedures for the same set of location data collected on the same spatial scale, 'location' ambiguity is inevitably created (Shavit and Griesemer 2009). Furthermore, when performing an individual measurement in the smallest spatial scale of that study, a) the measurement device had to be typically relocated to different longitudinal and latitudinal coordinates when positioned according to different concepts of space (even in the uncommon incidents when the device maintained its lat./long. coordinates, its location was empirically different as its description as a 'location');[5] and b) a barrier for communication was clear, translation was lost, and decisions were based upon hierarchy[6] or complete separation of the data[7]. On a later

[5] For example, there were two different maps of the Harvard Forest – with and without the location of each tree – which were deliberately kept separate. Choosing a location was made by randomly selecting a block on the blank map, yet when positioning a trap in the field, one repeatedly had to change its position because of the trees.

[6] At the MVZ resurvey, the lead researcher in the field acknowledged the dissatisfaction of his colleague from his interactionist space. Observation on August 25, 2007. In the Harvard Forest, the lead researcher decided on the locations beforehand and the traps were constantly maintained - interview from May 27, 2010. In that sense, there were no independent revisits so it is unclear if there was replication or one very long survey.

[7] The same applied to the Israeli resurveys: observation on May 16, 2005, June 6-7 2005 and September 3, 2008, interviews on April 23, 2009 and July 6, 2009. At Harvard Forest, observations and interviews conducted on May 26-28, 2010; June 6-8, 2012, July 30, 2012.

reflection, the researchers did not say that the work procedures used by their colleagues required more time or effort[8] or that the statements delivered by their collaborators were false, but rather: "it did not make any sense"[9], or "he is smart, I simply could not understand why he was so stubborn on this issue"[10], and often they only smiled gently and said: "I'm sorry, I could not do it the way they [or: he] wanted it".[11]

This clear-cut empirical gap, however, within all these biodiversity studies, do not lead to any disagreements on the overall answer to the question on species location and species distribution, since that answer was provided by aggregating results across higher spatial scales – that is, the plot/s or transact/s that encompass multiple individual measurements – and it was easily agreed that there are incompatible ways to validate/select and analyze that aggregated data. Acknowledging the barrier at the individual trap made

[8] Although one can interpret what the MVZ scientist said: "it would have been a total waste of time" (August 25, 2007) as a strictly practical or heuristic criticism, I understood it as a criticism on meaning, a precursor to the follow-up sentence: "it just made no practical sense!" (Ibid.)

[9] Interview with MVZ scientist March 23, 2007.

[10] Interview with Israeli scientists on April 23, 2009.

[11] Interview with Israeli scientists on July 6, 2009, February 9, 2010, and MVZ scientists on March 23, 2007 and August 25, 2007. For an exact citation: only one of the two Israelis used the word "sorry" and both MVZ scientists said "he" rather than "they".

scientist frequently alternate between spatial scales as "the relevant smallest scale".[12] They juxtaposed different concepts of space rather than seek a single concept for all levels (Shavit and Greisemer 2009), which facilitated the emergence of a productive scientific discourse (Shavit 2013;[13] Shavit and Ellison 2014, accepted for publication) and for different models to be recognized as useful alternatives (Shavit and Griesemer 2011b).

The philosophical literature on the concepts of incommensurability (Kuhn 1962; Feyerabend 1962), underdetermination (Duhem 1969; Quine 1953) and indeterminacy of translation (Quine 1960, 1990) seems especially relevant in this case. For both incommensurability and indeterminacy of translation "the paradoxical situation stems from meaning variance – the same terms have different meanings in the seemingly incompatible theories" (Ben Menahem 2006, 11), yet only "incommensurability implies that from the perspective of one paradigm (theory), the alternative is not simply false, but makes no sense at all" (Ibid). Listening to biologists debate, frequently surveying the same location

[12] The relevant smallest spatial scale for the theoretical MVZ ecologist, was the average of a transact line with 50 traps, while for the collector, it was the individual trap on that line (September 4, 2006); the smallest relevant scale for the Hamaarag was a single trap for the hierarchical sampling and a patch-type with three such traps for the landscape sampling (February 7, 2009) and at Harvard Forest, the smallest relevant scale was the experimental block with multiple traps (June 8, 2012).

[13] During a follow up symposiums on April 18, 2013 in Jerusalem (Israel), and on August 8, 2014 in Minneapolis (Minnesota), museum collectors, experimental ecologists and bioinformatics discussed their mutual problems of replication.

creates the impression that this is indeed a clear case of incommensurability.

In the next section, I will employ these concepts for describing a basic problem of replication in a manner that makes the scientists' disagreement more sensible than bizarre, which is presumably a better description.[14] In addition, taking note of the routine details of the scientific practice would clarify a common philosophical conflation between incommensurability and empirical equivalence (Ben Menahem 1990, 2006) and should therefore assist in avoiding it. That is, philosophical involvement in the routine scientific work not only helps to better describe and understand science – the standard role for the philosophy *of* science – but may also illustrate the benefits of philosophy *for* science (Griesemer 2011), at least for some scientists and philosophers of science.

3) Indeterminacy and Incommensurability

Philosophical discourse is replete with conflation of 'incommensurability' with 'empirical equivalence' (Ben Menahem, 1990). Given the time and space constraints, I will not address the longstanding controversy over 'incommensurability', the more recent debates over its history (Agassi 2002; Oberheim 2005), or its compatibility with scientific realism or progress (Demir 2008; Davis 2013).

Instead, I will rely on Ben Menahem's (1990, 2006) successful disconnection of a particular conflation regarding 'incommensurability': its association with empirical equivalence between semantically non-equivalent theories, and, as a result, the conflation of 'incommensurability' with 'indeterminacy of translation' and the common phrase "no fact of the matter". The blame could be placed on Kuhn's 1962/2005 explicit claim that

[14] To rationalize these assumptions, see Quine's 1960 and Davidson's 1984 "Principle of charity."

paradigms are incommensurable – i.e. not inter-translatable (Kuhn 1990) – and are *therefore* equivalent in the sense that there is "no fact of the matter" as to which paradigm to adopt (Ibid.194). However, as Ben Menahem (2006) had demonstrated, this conclusion does not follow, nor does it conform to other well-known examples of equivalence (for example Poincaré's argument for the empirical equivalence of different geometries (Ibid. Ch. 2). Briefly stated, incommensurability and indeterminacy are *not* closely related.

Then what is 'indeterminacy of translation'? "The thesis is then this: manuals for translating one language into another can be set up in divergent ways, all compatible with the totality of speech dispositions, yet incompatible with one another" (Quine 2004, 120). There is *no barrier* of communication. However, due to the lack of logical inference from observational to theoretical sentences – i.e. the underdetermination of theories – very different sentences can fit the same observation sentence rather than a one-to-one relationship between theory and data. That is, the 'meaning' of the data is not a determined entity that is somehow "captured in our minds" and is independent from its translation.

In our case studies, a "repeatable survey to the same location" – i.e. the practical meaning in terms of a detailed protocol in recording the location of our measurement – was constantly translated. Such explicit discussions between adherents of the different concepts of 'location' – prioritizing different practices at different spatial scales – occurred when theoretical considerations of diverge statistical packages, based on drawing aggregated results from higher spatial scales, came to the forefront. Although the exogenous and interactionist concepts of space were evaluated differently by various researchers and cultural-research bodies, and unlike the breakdown of communication outdoors at the

single trap scale, in all of the three cases at the lab, researchers could agree on a manual of translating this concept into routine practices to ensure “valid replication” and “high quality data”.

Quine distinguishes the indeterminacy of translation, which is manifested when different (and incompatible) sentences correlate to the same empirical data from the indeterminacy of reference (what he terms "ontological relativity"). A reference is indeterminate when the terms of the *same* sentence (or theory) could be correlated with the world in different ways (Quine, 1990). In this case, different empirical content may fit the same theatrical sentence, and there is no fact of the matter. In all of our case studies, such indeterminacy of reference have occurred, and researchers could easily agree on the truth value of sentences describing the survey results, as they were alternating between their incompatible interpretations. For example, researchers agreed on the number of organisms and species detected on a transact line, as they were alternating between models that hold incompatible assumptions on the causes of species detection on that transact. To test if a new survey of species' occupancy repeats the old survey well enough, one also has to model the variance in detection. Even if thirty *Peromyscus maniculatus* (deer mouse) were actually taken from Yosemite Valley in 2013 and 1913 - whether or not this truthful sentence implies that the deer mouse population did not change after spending a century in Yosemite, depends on the modeling of the variance of the collectors’ detection efforts, method era, or other parameters. Different environments – where the deer mouse population grows, declines, or unchanged– can correlate the same result, and researchers maintained their skepticism on the ontological interpretation of their models.

Such indeterminacy of reference is differentiated from indeterminacy of translation, as the former is involved with interpreting model results and the latter with outlining a protocol, yet they both relate to validation of survey repeatability and both are clearly different from the incommensurability of 'location' mentioned earlier.

4) Conclusion

In this article, I argued for a closer look at the seemingly mundane concepts of replication and location, by unfolding their divergent meanings, conceptual interties and impact on longstanding confusions in the philosophy of science. On that note, the common conflation between 'incommensurability' and 'empirical equivalence', 'indeterminacy of translation', 'indeterminacy of reference', 'incommensurability', and 'indeterminacy of reference', were easily and forthrightly avoided when taking note of the very different behaviors and procedures biologists use in the context of their work when re-surveying the same location. To clarify them and perhaps other controversial concepts, involvement in the scientific work seem to alleviate philosophical confusion. Observing the manner in which biologists use the terms 'location' and 'replication' revealed new distinctions on two different concepts of space – exogenous and interactionist – which adhere to different working standards without a common measurement, and to three different senses of replication – repeatability, reproducibility and retrieval – used differently at various stages of their work. Observing the scientists enabled the philosopher to raise new questions, find new conceptual distinctions and problematize the obvious. The benefits of this approach should not surprise us. After all, it has already been said that: " 'To give a new concept' can only mean to introduce a new deployment of a concept, a new practice" (Wittgenstein,

1978, 432). The aim of such a dialog between a practically involved philosopher and reflective scientists is not to transform either of which, but to build disciplinary bridges while closely minding the gaps between them.

References

Agassi, Joseph, 2002. “Kuhn’s Way." Philosophy of the Social Sciences 32: 394-430.

Barad, Karen. "Meeting the Universe Halfway, Quantum Physics and the Entanglement of Matter and Meaning." Durham: Duke University Press.

Ben Menahem, Yemima. 1990. “Equivalent Descriptions.” The British Journal of the Philosophy of Science 41: 261-79.

---------- 2006. "Conventionalism." Cambridge: Cambridge University Press.

Black, Jeremy. 1997. "Maps and History, Constructing Images of the Past." New Haven: Yale University Press.

Bissell, Mina. 2013. “The Risks of the Replication Drive. Nature 503: 333-34.

Bowker, Geoffrey C. 2005. "Memory Practices in the Sciences." Cambridge Massachusetts: M.I.T. Press.

Boyle, Robert. 1660. "New Experiments Physico-Mechanical: Touching the Spring of Air and their Effects." In "The Works of Thomas Boyle," edited by Mina Hunter and E. B. Davis. Vol. 1: 116-323. London UK: Pikering and Chatto 1999.

Cassey Phillip and Tom M. Blackburn. 2006. "Reproducibility and Repeatability in ecology." BioScience 56: 958-59.

Code, Lorraine. 2006. "Ecological Thinking: The politics of Epistemic Location." Oxford: Oxford University Press.

Collins, Harry. M. 1992. "Changing Order, Replication and Induction in Scientific Practice." Chicago: University of Chicago Press.

Davidson, Donald. 1984. "On the Very Idea of a Conceptual Scheme" in *Inquires into Truth and Interpretation*, Oxford: Clarendon Press: 183-198.

Davis, Alex. 2013. “Kuhn on Incommensurability and Theory Choice.” Studies in the History and Philosophy of Science 44: 571-79.

Duhem, Pierre. 1969. "To Save the Phenomena." Translated by E. Doland and C. Maschler. Chicago: Chicago University Press.

Ellison, Aaron. M., Leon, J. Osterweil. Hadley, A. Wise, Emery. Boose, L. Clarke, David. R. Foster, et al. 2006. "An Analytic Web to Support the Analysis and Synthesis of Ecological Data." Ecology 87: 1354-58.

Ellison, Aaron. M., Audrey Barker-Plotkin, David. R. Foster, and David A. Orwig. 2010. "Experimentally Testing the Role of Foundation Species in Forests: The Harvard Forest Hemlock Removal Experiment." Methods in Ecology and Evolution 1: 168-79.

Feyerabend, Paul 1962. "Explanation, Reduction and Empiricism." Feigl and Maxwell, 28-97.

Gerson, Elihu. 1998. "The American System of Research: Evolutionary Biology 1890-1950″ PhD. Dissertation Department of Social Studies, Chicago: The University of Chicago.

---------- 2007. "Reach, Bracket, and the Limits of Rationalized Coordination: Some Challenges for CSCW." Mark S. Ackerman et al. (eds.), Resources, Co-Evolution and Artifacts: Theory in CSCW. Dordrecht: Springer 193-220.

Griesemer, James R. 2011. “Philosophy and Tinkering.” Review of William C. Wimsatt. "Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.” Biology and Philosophy 26: 269-79.

---------- 2014. "Reproduction and the Scaffolded Development of Hybrids," Developing Scaffolds in Evolution, Culture, and Cognition. Caporael et al.(eds.) Cambridge, MA: MIT Press: 23-55.

Griffiths Paul .E. and Russell D. Gray. 1994. "Developmental Systems and Evolutionary Explanations." Journal of Philosophy 91: 277-304.

Guisan, Antoine. and Wilfred Thuiller. 2005. "Predicting Species Distribution: Offering More than Simple Habitat Models.” Ecology Letters 8: 993-1009.

Gugerli, David. 1998. "Politics on the Topographer's Table: The Helvetic Triangulation of Cartography, Politics, and Representation." Inscribing Science, Timothy Lenoir (ed.), Stanford: Stanford University Press 91-118.

Gunn, William. 2014. "Nature’s Correspondence." Nature 505: 26.

Hayden, E. C. 2013. "Weak Statistical Standards Implicated in Scientific Irreproducibility." Nature News 10.1038/nature: 14131

Hacking, Ian 1983. "Representing and Intervening." Cambridge: Cambridge University Press.

Hadot, Pierre. 1995. "What is Ancient Philosophy?" Translation: M. Chase. Cambridge MA: Harvard University Press 2002.

"IEEE Standard Computer Dictionary." A Compilation of IEEE Standard Computer Glossaries. New York: Institute of Electrical and Electronics Engineers, 1990.

Demir, Ipek. 2008. “Incomensurabilities in the Work of Thomas Kuhn" Studies in the History and Philosophy of Science 29: 133-42.

Jones, J., G. Lambrinos, Theresa Sinicrope Talley, and W.G.Willson. 2007. “Ecosystem engineering in space and time.” Ecology Letters 10: 153-64.

Kohler, Robert E. 2006. "All Creatures, Naturalists, Collectors and Biodiversity 1850-1950." Princeton: Princeton University Press, 2006.

----------"Practice and Place in Twentieth Century Field Biology: A Comment." 2012. Journal of the History of Biology 45: 579-586.

Kuhn, Thomas. S. 1962/1970 "The Structure of Scientific Revolutions." Second edition. Chicago: The University of Chicago Press.

Latour, Bruno. 1987. "Science in Action: How to Follow Scientists and Engineers Through Society." Cambridge MA: Harvard University Press.

Levins R. and R. Lewontin. 1985. "The Dialectical Biologist.” Cambridge MA: Harvard University Press.

Lloyd, Elizabeth. A. 2010. "Confirmation and Robustness of Climate Models." Philosophy of Science 77: 971-84.

Longino, Helen E. 2004. "How Values Can Be Good for Science," Science, Values and Objectivity. P. Machamer and Gottfried Wolters (eds.), Pittsburg: University of Pittsburg Press: 127-42.

Oberheim, Erik. 2005. “On the historical Origins of the Contemporary Notion of Incommensurability.” Studies in the History and Philosophy of Science 36: 363-90.

Popper, Karl R. 1959. [1992]. "The logic of Scientific Discovery." New York: Rutledge.

Michener, William. K. and Michelle. B. Jones. 2012. "Econiformatics: Supporting Ecology as a Data-Intensive Science." Trends in Ecology and Evolution 27: 85-92.

Odling-Smee, John F., Kevin N. Laland, and Michael.W. Feldman. 2003."Niche Construction: The Neglected Process in Evolution." Princeton: Princeton University Press.

Oyama, Susan. 1985. "The Ontogeny of Information: Developmental Systems and Evolution. Durham: Duke University Press, 2000.

Quine, William Van Orman. 1953."Two Dogmas of Empiricism" in *From a Logical Point of View*. Cambridge MA. Harvard University Press: 20-46.

---------- 1960. "Word and Object." Cambridge MA: Harvard University Press.

----------1990. “Three Indeterminacies” in Barrett and Gibson (eds.) Perspectives on Quine, Oxford: B. Blackwell, 11-16.

---------- 2004. "Quintessence, Basic Readings from the Philosophy of WilliamVan Orman Quine,” Ronald F. Gibson (ed.). Cambridge: Harvard University Press.

Redi, F.1668. "Experiments on the Generation of Insects." Translated by M. Bigelow. Chicago IL: Open Court 1909.

Robinson, T. M. 1987. “*Heraclitus*’s Fragments.” Toronto: University of Toronto Press.

Russell, J. F. 2013. "If a Job is Worth Doing, it is Worth Doing Twice." Nature 296: 7443.

Sanderson, Stephen K. 2013. "Bloggers Put Chemical Reactions through the Replication Mill." Nature News. 10.1038/nature:12262.

Shrader-Frechette, Kristin.S., and E.D. McCoy. (1993) "Methods in Ecology: Strategies for Conservation." Cambridge: Cambridge University Press.

Shiva, Vandana. 2000. "Tomorrow's Biodiversity." New York; Thames &Hudson.

Shapin, Steven and Simon Schaffer. 1985. "Leviathan and the Air-Pump Hobbes, Boyle, and the Experimental Life." Princeton: Princeton University Press.

Shavit, Ayelet and James R. Griesemer. 2009. "There and Back, or, the Problem of Locality in Biodiversity Research." Philosophy of Science 76: 273-24.

---------- "Transforming Objects into Data: How Minute Technicalities of Recording Species Location Entrench a Basic Theoretical Challenge for Biodiversity." 2011a, in *Science in the Context of Application*, edited by M. Carrier and A. Nordmann, 169-93. Bielefeld, Germany: Zentrum für interdisziplinäre Forschung (ZiF), 2011a.

———-"Mind the Gaps: Why are Niche Construction Processes so Rarely used?" 2011b, in *Lamarckian Transformations*, edited by S. Gisis and E. Jablonka, 307-17. Cambridge Mass: MIT press

Shavit, Ayelet 2013. "Replication Standards in long-term Research, Integrating the Field and Database Perspectives for Future Management." Ecological Society of America Bulletin 94: 395-397.

Shavit, Ayelet. and Aaron. R. Ellison (eds.). 2014. "Stepping in the Same River Twice." New Haven: Yale University Press. Accepted for publication.

Sterelny, Kim. 2001. “Niche Construction, Developmental Systems, and the Extended Replicator” in C*ycles of Contingency; Developmental Systems and Evolution.* Oyama, P.E. Griffiths, and R.D. Gray (eds.) Cambridge, MA: The MIT Press: 333-49.

Strasser, Bruno J. 2008. "GenBank - Natural History in the 21st Century?" in *Science* 32: 537-38.

Tingley, Morgan W. and Steven R. Beissinger. 2009. "Detecting Range Shifts from Historical Species Occurrences: New Perspectives on Old Data" in *Trends in Ecology and Evolution* 24: 625-33.

Wimsatt, William C. 2007. "Re-Engineering Philosophy for Limited Beings." Cambridge, Massachusetts: Harvard University Press.

Willson, Edward Osborne. 1992. "The Diversity of Life." Cambridge MA: Harvard University Press.

Wittgenstein, Ludwig. 1953. Philosophical Investigations, translation: G.E.M.Anscombe, eds. Gertrude Elizabeth Margaret Anscombe, and Rush Rhees. Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1956/1983 "Remarks on the Foundations of Mathematics." George Hhenrik von Wright and Rush Rhees (eds.), Gertrude Elizabeth Mmargaret Anscombe (translation). Cambridge MA: The MIT press.

Psychiatric Progress and The Assumption of Diagnostic Discrimination

Kathryn Tabb

Abstract: The failure of psychiatry to validate its diagnostic constructs is often attributed to the prioritizing of reliability over validity in the structure and content of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM). Here I argue that in fact what has retarded biomedical approaches to psychopathology is unwarranted optimism about *diagnostic discrimination*: the assumption that our diagnostic tests group patients together in ways that allow for relevant facts about mental disorder to be discovered. I consider the Research Domain Criteria (RDoC) framework as a new paradigm for classifying objects of psychiatric research that solves some of the challenges brought on by this assumption.

1. Introduction

The architects of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (1980), a task force of the American Psychiatric Association (APA), are often held to have sacrificed validity for reliability in constructing the manual's categories (Sadler, Hulgus, and Agich 1994; Kendell and Jablensky 2003; Andreasen 2007). According to this view, the DSM went wrong when it adopted an operationalist stance focusing on atheoretical observational criteria, an ecumenical approach that made it easier to apply diagnoses consistently across practitioners and contexts. Without an understanding of etiology, the argument continues, the real contours of psychopathology have not been demarcated, and psychiatry has not been able to identify disease entities akin to those in the rest of medicine (Murphy 2006). This narrative implies that psychiatrists incorporated the operationalism of the DSM into their research methodology, and were accordingly inhibited or uninterested in the pursuit of causal explanations. The solution to psychiatry's validity crisis, it has been suggested, is to refocus psychiatric research on causal mechanisms (Murphy 2006; Kendler 2011; Kendler et al. 2010).

Here I argue that the DSM stands in the way of valid diagnostic categories not merely because it codifies test criteria that have not been validated, but also on account of its role in the research setting. Due to its widespread use in the framing of scientific hypotheses about mental disorder, the manual plays a central role in shaping the objects and methods of psychiatric inquiry. In particular, its diagnostic criteria are widely used to gather test populations for psychiatric studies. When the DSM is employed in this way, the implicit assumption is that the criteria for diagnosing clinical types can also successfully pick out populations about which relevant biomedical facts can be discovered. I will refer to this as the *assumption of diagnostic discrimination.* This assumption is only justified if there is reason to believe that patients meeting diagnostic criteria for a given disorder also share one or more experiential, neurological, genetic, or other abnormalities.

The first aim of this paper (constituting Section 2) is to make explicit the role of the assumption of diagnostic discrimination in psychiatric research, specifically when that research uses DSM criteria to gather test populations. I show that the assumption is implicitly rejected in the Research Domain Criteria (RDoC) project, a new classification tool for psychiatric researchers introduced by the National Institute for Mental Health (NIMH). My second aim is to argue a pessimistic view of diagnostic discrimination, on historical and methodological grounds (Section 3). Finally I consider possible rebuttals for three of my claims: that there are no *a priori* grounds for optimism about diagnostic discrimination; that an alternative classification method would mitigate its risks; and that the assumption is ultimately such a bad one for researchers to make.

2. What is the assumption of diagnostic discrimination?

While the absence of valid categories in psychiatry is often noted, there is little consensus about what the term "validity" means in the psychiatric context. Olbert (unpublished) has identified

Tabb 3

fourteen distinct uses of the term in the literature, and suggests that there are significantly more in operation. The usage of the term in psychometrics, from which the psychiatric usage has been developed, is no less fraught; one study identified one hundred and twenty-two different subtypes of validity (Newton and Shaw 2013). The term was originally applied to psychological tests, and was used to calibrate how well a test measured what it was intended to measure. Validity was originally evaluated through the correlation of test scores with other criteria, such as alternative test outcomes. A list of criteria that could establish the validity of these inferences about psychiatric kinds was introduced in Robins and Guze (1970) and updated in Kendler (1980). Andreasen (1995) introduced a "second structural program for the validation of psychiatric diagnosis" which incorporated validators from neuroscience, genetics, and the biomedical sciences. Such validators range from characteristic course and family aggregation to genetic abnormalities and neural mechanisms.

Since diagnostic categories can be said to be measurement instruments in only a loosely analogous sense (Blashfield and Livesley 1991), psychiatrists tend to speak of validity instead as an attribute of the inferences made through diagnosis about purported disease entities. Here I follow psychiatrists themselves in employing the term "valid" to refer to psychiatric constructs that "approximate reality." Under the dominant biomedical paradigm in psychiatry, a valid diagnostic construct is one that categorizes patients who all share the same underlying physiological dysfunction. Critics of the DSM point out that none of the manual's categories have yet been validated in this sense, in so far as no account of a complete causal pathway to a mental disorder has been empirically demonstrated (Kapur et al. 2012).

My question here does not concern whether psychiatric kinds are valid, but rather whether the categories of the DSM, when used as instruments to collect test populations for research purposes, successfully congregate patients about whom relevant facts can be gathered. Optimists

about this question are often committed to what I term the *assumption of diagnostic discrimination*, that is, *the assumption that our diagnostic tests[1] group patients together in ways that allow for relevant facts about mental disorder to be discovered.* For the purposes of this discussion, *relevant* facts are those about the underlying mechanisms causing the signs and symptoms with which patients present. They are the sorts of facts that psychiatric researchers working in the biomedical sciences hope to find: genetic signatures, neurological or cognitive dysfunctions, focal brain lesions, and so forth. Diagnostic discrimination may be a more or less justified assumption for those interested in other sorts of inferences, as I will consider briefly in section 5.1.

I borrow the term "diagnostic discrimination" from psychometrics, where it is defined as the statistical assessment of how a diagnostic test compares with a gold standard, measured by the test's specificity, sensitivity, predictive value, and likelihood ratios (Knottnerus and Buntix 2009, 4). Discrimination in this sense is inapplicable in psychiatry, which lacks any authoritative tests that would allow for the assessment of the sensitivity or specificity of the DSM's categories. In my argument it is invoked as an aspirational term, signifying an ideal rather than a measure. I am interested in the particular epistemic stance that evinces optimism about whether our diagnostic categories effectively group together patients homogeneous for the real objects of interest for biomedical psychiatry. The extent the DSM's criteria are discriminative for the purposes of biomedical research is, of course, an empirical question, and diagnostic discrimination will surely vary across the manual's constructs. My aim is not primarily to offer any empirical assessments, but rather to raise some concerns about the warrant for *prima facie* optimism about discrimination.

[1]By "tests" I refer to either the diagnostic criteria of the DSM itself or diagnostic screens based on these criteria. Obviously diagnostic discrimination could be proposed about other diagnostic methods (e.g., the *Psychodynamic Diagnostic Manual*) but my focus here is on the DSM.

Unfortunately, careful attention to the problem of validity is entirely compatible with a naive commitment to diagnostic discrimination. Study designs that use DSM criteria to select research samples may assume that those samples will be homogeneous for certain sorts of pathogenic mechanisms.[2] Even those profoundly dissatisfied with DSM categories may employ its criteria in order to locate latent constructs they hope to use to revise and perfect the manual. The DSM's central role in the research context, specifically in guiding the selection of test populations and establishing targets for explanation, is not only entrenched by historical precedent but also held firm by the hand of the biomedical marketplace; funding bodies have traditionally preferred research that is directly pertinent to perceived clinical needs. This has led to a focus on the iterative validation of diagnostic constructs, especially the search for the causal mechanisms that can undergird new therapies.

In the following section I explore the role of diagnostic discrimination in the history of psychiatric research, and suggest that this history should lead us to be pessimistic about the assumption's warrant. In my fourth section I will make the conceptual case against optimism about whether our diagnoses are discriminative, and consider an alternative tool for gathering test populations that does not rely on this risky assumption.

3. The case for pessimism: a historical argument

A valid taxonomy has historically been viewed as the first step in psychiatric research. Influential theorists of psychiatric validity have imagined a boot-strapping model, in which the first phase of achieving validity involves settling on a clinical description of diagnostic kinds (Kendell and

[2] Imagine a psychopharmaceutical study with a simple design in which drug response is tested in a clinical population of subjects sharing a diagnosis. If the assumption of diagnostic discrimination is in play, a 60% response rate will be interpreted as demonstrating that the drug is effective 60% of the time. Once the assumption is questioned, alternative hypotheses—such as that 60% of patients sharing a diagnosis share a specific underlying mechanism affected (with 100% efficacy) by the drug—can be considered.

Jablensky 2003). Andreasen, for example, writes that only "once a reliable method is applied to define symptoms or delineate a potential diagnostic category or dimension of psychopathology" can "these variables then be validated by examining their relationship to external measures" (1995, 162). The DSMs have, historically, provided the independent variable for studies attempting to validate psychiatric kinds.

However, the origins of today's diagnostic categories do not offer confidence that they will be discriminative in the relevant way. Despite the ideal of a scientifically objective system, psychiatric kinds are historically embedded concepts, traceable to different strata of the discipline's past. The aim of the first edition of the DSM, published in 1952, was to collect statistical information. Throughout the history of the manual, ambitious task forces have attempted to revise the DSM's categories on the basis of contemporary methods and knowledge, rather than in the terms of decades-old census projects and nineteenth-century theory. With somatic medicine as the benchmark, discriminative diagnoses were considered the ideal targets for validation by early advocates of the medical model in psychiatry (Klerman 1978); the architects of the DSM-III prioritized the construction of diagnostic categories based on "distilled clinical research experience" as the "first and crucial taxonomic step" (Feighner et al. 1972, 57) towards identifying valid constructs.

While Feighner et al. sought to reground psychiatric nosology on empirical foundations, their criteria (which formed the template for the DSM-III) were in fact an amalgam of data and received clinical intuition, with many of the basic taxonomic divisions being inherited unchallenged (Kendler 2009). Similarly, the main architects for the most recent revision, the DSM-5, announced the need to "transcend the limitations of the current DSM paradigm" so that the new DSM could provide research criteria "not constrained by the requirements of the neo-Kraepelinian categorical approach currently adopted" (Kupfer, First, and Regier 2008, xxii).

Tabb 7

In the end, however, with some exceptions (such as the reconfiguration of subtypes of autism on a spectrum and the removal of subtypes for schizophrenia) the nosological structure remained relatively stable.

Since the DSM is primarily intended to serve a clinical population, it makes sense that latent constructs postulated but not demonstrated by biomedical researchers would be excluded. Until theories about underlying mechanisms can be correlated with signs and symptoms that present in the clinic (either behaviorally or as the result of testing), they are irrelevant to the task of diagnosing patients. As it stands, the biologization of psychiatric research has not led to the discovery of any laboratory markers for specific psychiatric conditions, and there remain no biological screens for psychopathology—only the checklists of the DSM itself, and the tests that are based on its operationalizations (Kapur 2012). Decades of research into psychiatric and behavioral genetics have failed to turn up genes specific to particular disorders (though the heritability of types of psychopathology has been demonstrated [Merikangas and Risch 2003]) or neurological markers (despite advances in our understanding of the neurological underpinnings of signs and symptoms [Gillihan and Parens 2011]). The pharmaceutical industry has capitalized on optimism about a one-to-one correspondence between diagnosis, condition, and treatment; notable here is the historic relabeling of treatments specific to symptoms (e.g., "tranquilizers") as treatments specific to purported disease entities (e.g., "antipsychotics"). In spite of this, the heterogeneity of diagnostic profiles is matched by the heterogeneity of patient response to treatment. Nearly all psychopharmaceutical interventions are nonspecific, and none come close to working for all patients sharing a diagnosis, which would allow the DSM to be redrawn along the lines of what Radden has called "drug cartography" (2003).

All in all, neither the history of the manual nor the current state of the art in biomedical psychiatry can support the assumption of diagnostic discrimination. In the next section I argue

that the structure of the DSM also gives reasons for pessimism, drawing on criticisms made by a growing number of psychiatric researchers that their disappointing failure to validate the DSM's constructs is due to the fact that there is nothing for them to validate. Or, to put these judgments about the ontology of psychiatric kinds in my own epistemological terms: the diagnostic tests for psychiatric constructs are not discriminative in the relevant sense, in so far as little of interest from the perspective of biomedicine can be discovered about patients sharing a diagnosis beyond the recognition that they all present with (some of) the very signs and symptoms that constitute their diagnosis.

4. The case for pessimism: a conceptual argument

The first thing to be noted about the DSM's structure is that if etiopathogenic facts about mental disorders are forthcoming, they will not stand in simple causal relationships to the signs and symptoms that act as diagnostic criteria. As of its third edition the DSM's categories have been polythetic, requiring patients to present with only *n* symptoms out of a longer list in order to meet the threshold for a given disorder. The diversity of patients within each class is increased further because screens for psychopathology tend to have low thresholds, since the cost of a false-negative (abandoning a patient in need of care) is viewed as higher than a false-positive (giving unneeded treatment) (Ross 2014). This has allowed diagnostic criteria to cast wider nets, and for reliability to be improved. But as a result, the DSM's criteria allow for incredible diversity. For example, the DSM-5 permits patients to be diagnosed with post-traumatic stress disorder if they present with any one of 636,120 possible combinations of symptoms (Olbert et al., 2014). This may mean that patients sharing diagnosis have a range of underlying pathologies that cause these related but distinct manifestations. Relevant facts will explain this diversity either by revealing homogeneity beneath promiscuous clinical descriptions, or by ultimately arriving at disjunctive accounts of the mechanisms that undergird them. The likelihood of the former across psychiatric

diagnoses is doubtful, given the relative rarity of single causes underlying distinct clinical presentations in somatic medicine (Olbert, unpublished). In cases of the latter, the heterogeneity of conspecifics that make up DSM-derived research samples could hamper progress towards a discovery of these diverse mechanisms.

Some amount of symptomatic variation is frequently found among patients sharing a diagnosis in other types of disease, such as cancer or lupus, so heterogeneity on its own does not prove that the DSM's diagnoses are not discriminative. But the lack of compelling confirmations of psychiatry's taxonomic boundaries by genetics, epidemiology, neurophysiology, and other allied sciences is worrying, raising the question of whether the manual is useful for anything more than identifying phenotypic clusters (Meehl 1986). Turning to the DSM's use in the research setting, initial hopes that "zones of rarity" among diagnoses would emerge through the discovery of underlying mechanisms have not yet been fulfilled (Kendell and Jablensky 2003). Taxometric and epidemiological studies reveal that the enormous heterogeneity in symptoms and course actually contain recognizable sub-types that appear more frequently than others; however, underlying differences in causal pathways or mechanisms that could explain these trends have not been found (Nandi, Beard, and Galea 2009).

Recently, a new round of critics has suggested that the heterogeneity of test populations collected on the basis of DSM diagnostic criteria undermines these sorts of discoveries in psychopathology. Some believe that the best response would be to do away with diagnostic constructs as targets for validation (Hyman and Fenton 2003; Merikangas and Risch 2003). Their view is that explanations that facilitate intervention and recovery are better found at other levels—for example, the level of the symptom, the gene, or the neural mechanism. Sanislow et al. have written that "dependence on conventional nosologies leaves the enterprise of understanding mechanisms of psychopathology in the awkward position of assuming the validity of single

disorders and organizing research accordingly" (Sanislow et al. 2010, 2). In fact, validity is not assumed in such cases—the soundness of inferences about the diagnostic construct is not taken for granted, but rather is the object of investigation. What Sanislow et al. are reacting to is the assumption of diagnostic discrimination—the assumption that populations delineated by DSM categories are ripe for validation according to current biomedical standards.

This line of criticism is a reaction to cases like that described by Steven Hyman who, as the director of the NIMH in the late 1990s, became aware of and increasingly frustrated by the lack of research into treatments for the cognitive deficits of schizophrenia, among the most difficult and damaging symptoms experienced by patients. Hyman describes realizing that the lack of interest in cognitive symptoms was due to the bottleneck put on research by the DSM's diagnostic criteria, since cognitive deficits were not included in the manual. "Given the status of the DSM-IV criteria as the community consensus," Hyman writes of that time, "the U.S. Food and Drug Administration (FDA) held that it could not, by itself, recognize the cognitive symptoms of schizophrenia as an indication for the development and approval of new treatments" (Hyman 2010, 157). Recently, the DSM-5 Task Force has justified the continued lack of inclusion of cognitive symptoms quite explicitly, on the grounds that "cognition may not be useful as a differential diagnosis tool."[3]

Hyman's worry is that a vicious cycle is produced by the role of the DSM in research, such that the exclusion of a symptom (like cognitive deficit) from the manual for clinical reasons leads to the suppression of precisely the kind of research that would make its saliency for psychiatric practice clear. With his colleagues at the NIMH, Hyman began to construct a classification system for *research* that would allow scientists to apply for funding from the Institute without structuring their studies around DSM categories. Under the Research Domain Criteria rubric,

[3] http://www.DSM5.org/ProposedRevision/Pages/proposedrevision.aspx?rid=411#.

psychiatric investigators present their experiments as targeting fundamental components of mental functioning (or "research domains") that are drawn from allied sciences, instead of using DSM constructs. Research domains contribute one axis to the matrix that the NIMH has proposed for organizing psychiatric research, which is sub-divided into more specific "constructs"—for example, "reward valuation," "performance monitoring," or "attachment formation and maintenance." The other axis is "units of analysis," ranging from "genes" to "behavior."[4]

By encouraging the funding of research that investigates certain research domains at certain units of analysis, the RDoC changes the targets of validation from "clinical endpoints that have remained unchanged for decades" (Hyman and Fenton 2003, 351) to any sort of phenomenon relevant to psychopathology that may be viewed either as an extreme on a spectrum of human variation or as a dysfunctional structure or process. What is at stake with this new approach is the longstanding contention that psychiatry's scientific targets are best located through the same classificatory tools as those deployed in clinical practice. Rather than seeking to replace the DSM as a diagnostic manual, RDoC works as a classification protocol for researchers. It aims to encourage a profound shift in the way research samples are conceived of and assembled. In some cases, the translational approaches encouraged by the NIMH require the study of mechanisms that cut across traditional diagnostic categories. Now, instead of relying on DSM categories to gather research populations, RDoC researchers may gather whatever populations are pertinent to their domain of interest.

This method facilitates the roundabouts researchers have always used to precisify generic diagnostic screens to meet their own needs (Meehl 1986; Kutschenko 2011a). Test populations

[4] The matrix also includes a column for "paradigms," which are not units of analysis but rather scientific methods, frameworks, or tasks that are of use in the study of a particular construct.

need not even manifest homogeneous psychopathological symptoms, and indeed one of the aims of RDoC is to allow for the inclusion of patients typically ignored in research because they fall into a "not otherwise specified" category, as well as patients who show signs of mental distress but are below the threshold for diagnosis. So, for example, a group researching fear circuitry (*construct of interest*: fear/acute threat; *domain*: negative valence systems; *unit*: circuits) might use as their test population patients seeking medical help for anxiety, regardless of whether they meet any specific diagnostic criteria.[5]

The RDoC project avoids the pitfall of prematurely assuming diagnostic discrimination, although, as I discuss in Section 5.3, it still relies on other types of discrimination that may be faulty. Of interest here is that in order to liberate psychiatric research from the constraints of an unhelpful taxonomy, the NIMH has placed its bets for discrimination of research targets beyond the pages of the DSM. Debates over which sorts of objects are most worthy of study may continue to be played out under the RDoC through the distribution of funding dollars, but these judgments will be constrained by current epistemological and methodological commitments rather than nosological tradition. In contrast, when the DSM is used to design experimental protocols and present them to funding bodies it can act as a bottleneck, restricting research that cross-cuts or challenges existing diagnostic boundaries and excluding innovative explanatory approaches. If the DSM's categories are discriminative in the relevant sense, such a narrowing of focus is a boon to research. If not, the DSM is analogous to the lamppost in the tale of the man who makes the mistake of looking for his keys where the light is, instead of where he lost them.

[5] This example is borrowed from the NIMH's online materials about the RDoC—see http://www.nimh.nih.gov/research-priorities/rdoc/nimh-research-domain-criteria-rdoc.shtml#toc_studies for the full example. Accessed 6/18/14.

Tabb 13

5. In defense of optimism

I have argued that the DSM may retard progress in psychiatry not merely by codifying and enforcing diagnoses that may not be valid, but also by limiting the abilities of researchers to make original valid inferences about the nature of psychiatric disorder.[6] This effect is due to the widely-held but, I have argued, unjustified assumption in psychiatry that the manual's categories are the appropriate grounds on which to draw test populations for research purposes. In this section I consider three possible objections to my argument. The first is that if warrant for belief in diagnostic discrimination cannot be found in the DSM's history or biomedical psychiatry's track record, it can be found in clinical practice. The second is that *some* assumptions about discrimination must be made, and that the bottlenecking effects that these assumptions have on progress are a necessary cost of doing science. The third is that by giving up on validating the DSM's categories, psychiatry would lose track of its true targets, making the assumption of diagnostic discrimination a prerequisite for psychiatric research.

5.1 The Clinical Case for Diagnostic Discrimination

It has been assumed that if clinicians are able to separate patients into discrete kinds based on their symptomology there is good reason to anticipate that scientific validators will ultimately reinforce these divisions (Robins and Guze 1970). However, it seems that many clinicians themselves do not believe that the DSM accurately taxonomizes their patients. Studies of the actual usage of the manual suggest that clinicians find it primarily helpful for securing treatment options, and mostly ignore its complex polythetic structure (First and Westen 2007). Practitioners engage in diagnostic "bracket creep" to tweak coverage benefits and duck the restrictions that

[6] There are, of course, countless other powerful bottlenecks on psychiatric progress, among them that the brain is far more complex than other medical objects and that explanations of psychopathology from a biomedical perspective may well always be (to a greater degree than elsewhere in medicine) incomplete without contributions from psychology, the social sciences, and even the humanities.

insurance companies put on their ability to utilize their expert judgment (Bowker and Star 1999). Ethnographic research reveals that diagnoses often follow *after* treatment, rather than guiding it (Whooley 2010, 461). If the manual's ubiquity in clinical practice is due to its integral role in the larger machinery of industrial and corporate healthcare, rather than its accurate representation of clinical types, any argument for diagnostic discrimination on these grounds is unsound.

Further evidence that the manual's diagnostic constructs do not accurately represent clinical concepts of disorder comes from the widespread alarm over the deprecation of the experience of the patient due to the DSM's reductive approach to description (Andreasen 2007). The DSM's operationalized descriptions neglect the fact that "in addition to manifesting the relatively direct consequences of neurobiological abnormalities," patients "react to their abnormalities in all kinds of ways that may sometimes require the categories of meaning and experience in order to be understood or explained" (Sass, Parnas, and Zahavi 2011, 16). Some phenomenologically-oriented clinicians and philosophers of psychiatry have suggested that these experiential aspects of mental illness that should themselves be targets for validation (Mishara and Schwartz 2010). Ipseity disturbance, for example, has been used to differentiate schizophrenia-spectrum disorders from other forms of psychosis (Henriksen and Parnas 2012; Parnas et al. 2005). Taken together, these criticisms suggest that the DSM categories do not reflect the clinical picture sufficiently to justify optimism about their utility in the research setting.

5.2 The inevitability of diagnostic discrimination

Another possible objection is that the assumption of discrimination is inevitable in psychiatric investigation, and that the DSM is not (uniquely) culpable. Studies dividing subjects into groups must be always depend on tests assumed to be discriminative for the construct in question. Strategies like RDoC, it could be argued, simply replace the diagnostic constructs of the DSM with other sorts of constructs, in this case the sub-categories of its proposed domains. The validity

of these constructs can surely also be challenged, and the organization of research methods and practices in accordance with them could also be restrictive.

My aim is not to dismiss the importance of discrimination in psychiatric research, nor to suggest that psychiatry can or should do without constructs altogether, but rather to challenge the assumption that the DSM's criteria are discriminative for research purposes. While the RDoC also relies on constructs, its architects have emphasized that these constructs are, first, completely open to revision and, second, explicitly designed to be broad enough to include the major paradigms *currently at play* within psychiatric research today. If the NIMH does not fulfill its promise to amend and expand the matrix's research domain criteria in accordance with shifts in the field, it could well end up with calcified categories that restrict research in the way that the DSM's categories have.

Notably, RDoC does not limit the conceivable *objects* of psychiatric research, which are not the same as the loci on the matrix at which the research falls. Rather than taxonomizing objects for psychiatric investigation, RDoC arranges domains of functioning in which such objects are located, providing for each a consensus definition and orienting researchers towards the available measures or elements across the units of analysis that could be used as variables for gathering populations for studies.[7] Accordingly, researchers have a significant amount of autonomy in the design of their research. As in all scientific research, their choice of construct and the tests they use to measure for it should be scrutinized closely by their peers.

5.3 The value of diagnostic kinds for psychiatric research

A final objection worth considering is whether giving up on diagnostic kinds is worth it—whether the gains to research productivity that would come from having discriminative targets have too high an epistemological or ethical cost. It can be argued that keeping psychiatry focused on

[7] http://grants.nih.gov/grants/guide/notice-files/NOT-MH-11-005.html

diagnostic kinds is the best way to avoid the reduction of the mentally ill to their component parts, which neglects the phenomenological core of psychopathology (McLaren 2011; Walter 2013). Thus there is a risk that the NIMH's own assumptions about the proper targets for psychiatric explanation may be crippling, potentially becoming (in Hyman's evocative term for the DSM) another "unintended epistemic prison" (Hyman 2010, 157).

The NIMH has made little secret of its preference for analysis at the level of brain circuits, based on the reasoning that it is at this level that science is most rapidly gaining insight into the underlying correlates of behavior (Insel et al. 2010). However, this approach has garnered accusations that the RDoC is "mindless" (Frances 2013), that is, symptomatic of "the profession's intent to complete its abandonment of the mind as the localization and source of our suffering" (Greenberg 2013, 342). In response Bolton (2013) has argued that the NIMH's claim that "all mental diseases are brain diseases" need not be reductionistic insofar as the brain can be seen as integrated into a complex network of causal relations that extend beyond the individual. Other advocates of the RDoC framework suggest it might give empirical grounding to psychotherapeutic as well as pharmotherapeutic approaches (Morris and Cuthbert 2012, 31). However, especially in light the NIMH's increasingly enthusiastic pursuit of basic science even as "fundamental and important questions regarding health services, psychosocial treatments, conceptual issues, public health, and patient initiatives remain marginally funded" (Sadler 2013, 29), it remains to be seen whether the NIMH will be truly ecumenical in the distribution of research dollars across the columns of their matrix.

The RDoC project's purported reductionism differs in an important way from the epistemic bottleneck of the DSM, however, insofar as it increases the conceptual and methodological distance between the laboratory and the clinic rather than collapsing it. If the pretense is abandoned that psychiatry's scientific and practical objects are one and the same, the fits and

starts of the NIMH's descriptive project need not immediately impact clinical nosology, nor need its reductive approach be directly imported into clinical practice. Solomon has argued that while expert disagreement can be generative in science, the value of stable consensus is higher in medicine, where the loss of epistemological authority can be dangerous (Solomon 2014). Her claims are vindicated by the widely expressed view that even the minor modifications of diagnostic categories found in each new edition of the DSM can be greatly harmful to patients (Frances 2009). As Schaffner has suggested, clinical research might continue to make progress on refining our understanding of psychopathology at "higher levels of aggregation" while projects facilitated by the RDoC framework work to reveal the complex and diverse "many-many relations" that make validity such a challenge (Schaffner 2012, 184). However, if the DSM stops playing its role as an epistemic hub (Kutschenko 2011b), the integration of psychiatric knowledge into therapeutics will need to be re-imagined—a project well beyond the scope of this paper.

6. Conclusion: Implications for Philosophy of Psychiatry

Diverse metaphysical orientations about the nature of the kindhood of diagnostic kinds are compatible with the assumption of diagnostic discrimination. Debates among philosophers of psychiatry over psychiatric kinds have focused on appraising these possible metaphysical stances, and there has recently been much effort to resolve the metaphysical nature of psychiatric kinds (Kincaid and Sullivan 2014). Insofar as the objects of diagnostic tests can be seen as either theoretical constructs or real entities, both realists and instrumentalists can beg the question of whether the DSM's diagnostic criteria are indeed discriminative. This project has distracted philosophers from the fact that optimism about the discrimination of the diagnostic criteria may not, in some or all cases, be warranted. We have no reason to doubt that diagnostic discrimination varies across the DSM's categories, rendering as ill formed the question of whether psychiatric kinds are natural, human, practical, constructed, etc. Since psychiatrists are

increasingly pursuing piecemeal causal explanations about constructs below the level of the diagnostic construct, they should follow Kincaid (2008) in leaving the question of diagnostic kindhood behind. Instead, philosophers can investigate the ways in which psychiatry stabilizes its diverse objects of research across disciplinary boundaries in the absence of the DSM's authoritative voice (Sullivan 2014).

References

Andreasen, Nancy. 1995. "The Validity of Psychiatric Diagnosis: New Models and Approaches." *America Journal of Psychiatry* 152:161–62.

Andreasen, Nancy. 2007. "DSM and the Death of Phenomenology in America: An Example of Unintended Consequences." *Schizophrenia Bulletin* 33 (1):108–12.

Bolton, Derek. 2013. "Should Mental Disorders Be Regarded as Brain Disorders? 21st Century Mental Health Sciences and Implications for Research and Training." *World Psychiatry* 12 (1): 24–25.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.

Feighner, John, Eli Robins, Samuel Guze, Robert A. Woodruff, George Winokur, and Rodrigo Muñoz. 1972. "Diagnostic Criteria for Use in Psychiatric Research." *Archives of General Psychiatry* 26:57–63.

First, Michael B., and Drew Westen. 2007. "Classification for Clinical Practice: How to Make ICD and DSM Better Able to Serve Clinicians." *International Review of Psychiatry* 19 (5):473–81.

Frances, Allen J. 2009. "A Warning Sign on the Road to DSM-V: Beware of Its Unintended Consequences." *Psychiatric Times* 26 (8):1–4.

Frances, Allen J. 2013. "The Role of Biological Tests in Psychiatric Diagnosis." *Huffington Post: Science*. July 25.

Gillihan, Seth J., and Erik Parens. 2011. "Should We Expect 'Neural Signatures' for DSM Diagnoses?" *The Journal of Clinical Psychiatry* 72 (10):1383–89.

Greenberg, Gary. 2014. *The Book of Woe: The DSM and the Unmaking of Psychiatry*. Blue Rider Press.

Hyman, Steven E. 2010. "The Diagnosis of Mental Disorders: The Problem of Reification." *Annual Review of Clinical Psychology* 6:155–79.

Hyman, Steven E., and Wayne S. Fenton. 2003. "What Are the Right Targets for Psychopharmacology?" *Science* January 17: 350–51.

Insel, Thomas R. 2013. "Transforming Diagnosis." Director's Blog, NIMH.

Kapur, Shitij, Anthony G. Phillips, and Thomas R. Insel. 2012. "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" *Molecular Psychiatry*, 17(12):1–6.

Kendler, Kenneth S. 1980. "The Nosologic Validity of Paranoia (Simple Delusional Disorder): A Review." *Archives of General Psychiatry* 37:699–706.

Kendler, Kenneth S., and R. A. Muñoz. 2010. "The Development of the Feighner Criteria: A Historical Perspective." *The American Journal of Psychiatry* 167:134–42.

Kendler, Kenneth S., Peter Zachar, and Carl Craver. 2009. "What Kinds of Things Are Psychiatric Disorders?" *Psychological Medicine* 41:1143–50.

Kincaid, Harold. 2008. "Do We Need Theory to Study Disease?" *Perspectives in Biology and Medicine* 51: 367–78.

Kincaid, Harold, and Jacqueline A. Sullivan, eds. 2014. *Classifying Psychopathology*. Cambridge, MA: MIT Press.

Knottnerus, J. A., and Frank Buntinx. 2009. *The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research.* West Sussex: Blackwell.

Kupfer, David J., Michael B. First, and Darrel A. Regier. 2008. *A Research Agenda for DSM-V*. Arlington: American Psychiatric Publishing.

Kutschenko, Lara K. 2011a. "In Quest of 'Good' Medical Classification Systems." *Medicine Studies*.

Kutschenko, Lara K. 2011b. "How to Make Sense of Broadly Applied Medical Classification Systems: Introducing Epistemic Hubs." *History and Philosophy of the Life Sciences* 33 (4):583–601.

McLaren, Niall. 2011. "Cells, Circuits, and Syndromes: a Critical Commentary on the NIMH Research Domain Criteria Project." *Ethical Human Psychology and Psychiatry* 13 (3):229–36.

Meehl, Paul E. 1986. "Diagnostic Taxa as Open Concepts: Metatheoretical and Statistical Questions About Reliability and Construct Validity in the Grand Strategy of Nosological Revision." In *Contemporary Directions in Psychopathology*, ed. Theodore Millon and Gerald L. Klerman, 215–31. New York: Guilford.

Merikangas, Kathleen Ries, and Neil Risch. 2003. "Will the Genomics Revolution Revolutionize Psychiatry?" *American Journal of Psychiatry* 160 (4): 625–35.

Nandi, Arijit, John R. Beard, and Sandro Galea. 2009. "Epidemiologic Heterogeneity of Common Mood and Anxiety Disorders Over the Lifecourse in the General Population: A Systematic Review." *BMC Psychiatry* 9:31.

Olbert, Charles (2014). *On the Mathematical Coherence of Psychiatric Diagnostic Categories: A Framework for Quantifying Heterogeneity Attributable to Polythetic Diagnostic Criteria*. Unpublished master's thesis, Fordham University, New York, NY.

Olbert, Charles, Gary J. Gala, and Larry A. Tupler. 2014. "Quantifying Heterogeneity Attributable to Polythetic Diagnostic Criteria: Theoretical Framework and Empirical Application." *Journal of Abnormal Psychology* 123 (2):452–62.

Parnas, Josef, Paul Møller, Tilo Kircher, Jørgen Thalbitzer, Lennart Jansson, Peter Handest, and Dan Zahavi. 2005. "EASE: Examination of Anomalous Self-Experience." *Psychopathology* 38 (5):236–58.

Poland, Jeffrey, and Barbara Von Eckardt. 2013. "Mapping the Domain of Mental Illness." In *Oxford Handbook of Philosophy and Psychiatry*, ed. K. W. M. Fulford, Martin Davies, Richard Gipps, George Graham, John Sadler, Giovanni Stanghellini, and Tim Thornton, 735–52. Oxford: Oxford University Press.

Radden, Jennifer. 2003. "Is This Dame Melancholy?: Equating Today's Depression and Past Melancholia." *Philosophy, Psychiatry & Psychology* 10(1): 37-52.

Robins, Eli, and Samuel B. Guze. 1970. "Establishment of Diagnostic Validity in Psychiatric Illness: Its Application to Schizophrenia." *American Journal of Psychiatry* 126 (7):983–87.

Ross, Don. 2014. "Syndrome Stabilization in Psychiatry: Pathological Gambling as a Case Study." In *Classifying Psychopathology*, ed. Harold Kincaid and Jacqueline A. Sullivan, 195–

208. Cambridge, MA: MIT Press.

Sadler, John Z. 2013. "Considering the Economy of DSM Alternatives." In *Making the DSM-5: Concepts and Controversies*, ed. Joel Paris and James Phillips, 21–38. New York: Springer.

Sanislow, Charles A., Daniel S. Pine, Kevin J. Quinn, Michael J. Kozak, Marjorie A. Garvey, Robert K. Heinssen, Philip Sung-En Wang, and Bruce N. Cuthbert. 2010. "Developing Constructs for Psychopathology Research: Research Domain Criteria." *Journal of Abnormal Psychology* 119 (4):631–39.

Sass, Louis A., Josef Parnas, and Dan Zahavi. 2011. "Phenomenological Psychopathology and Schizophrenia: Contemporary Approaches and Misunderstandings." *Philosophy Psychiatry and Psychology* 18 (1):1–23.

Schaffner, Kenneth F. 2012. A Philosophical Overview of the Problems of Validity for Psychiatric Disorders. In *Philosophical Issues in Psychiatry II: Nosology*, ed. Kenneth S. Kendler and Josef Parnas, 169–89. Oxford: Oxford University Press.

Solomon, Miriam. 2014. "Expert Disagreement and Medical Authority." In *Philosophical Issues in Psychiatry III: The Nature and Sources of Historical Change*, ed. Kenneth S. Kendler and Josef Parnas. Oxford: Oxford University Press.

Sullivan, Jacqueline A. 2014. "Stabilizing Mental Disorders: Prospects and Problems." In *Classifying Psychopathology*, ed. Harold Kincaid and Jacqueline A. Sullivan, 195–208. Cambridge MA: MIT Press.

Vaillant, George E. 1984. "The Disadvantages of DSM-III Outweigh Its Advantages." *American Journal of Psychiatry* 141 (4). American Psychiatric Association: 542–45.

Walter, Henrik. 2013. "The Third Wave of Biological Psychiatry." *Frontiers in Psychology* 4: Article 582.

Whooley, Owen. 2010. "Diagnostic Ambivalence: Psychiatric Workarounds and the Diagnostic and Statistical Manual of Mental Disorders." *Sociology of Health & Illness* 32 (3):452–69.

Confirmation Measures and Sensitivity

November 3, 2014

Abstract

Stevens (1946) draws a useful distinction between ordinal scales, interval scales, and ratio scales. Most recent discussions of confirmation measures have proceeded on the ordinal level of analysis. In this paper, I give a more quantitative analysis. In particular, I show that the requirement that our desired confirmation measure be at least an *interval* measure naturally yields necessary conditions that jointly entail the log-likelihood measure. Thus I conclude that the log-likelihood measure is the only good candidate interval measure.

1 Introduction

Suppose our preferred confirmation measure, c, outputs the numbers $c(H_1, E) = 0.1$, $c(H_2, E) = 0.2$, $c(H_3, E) = 0.3$, $c(H_4, E) = 50$ for hypotheses H_1, H_2, H_3, and H_4, given evidence E. It is natural to want to say that H_1 and H_2 are confirmed to roughly the same (low) degree by E, and that H_4 is confirmed by E to a much higher degree than either H_1 or H_2. We might also want to say that the *difference* in confirmation conferred by E on H_1 as opposed to on H_2 is the *same* as the difference in confirmation conferred by E on H_2 as opposed to on H_3. If we make any of the preceding assertions, we are implicitly relying on the assumption that it is legitimate to interpret the differences between the numbers outputted by measure c. In other words, we are assuming that c is at least an *interval measure* in the terminology of Stevens (1946). In this paper I will show how the preceding assumption, when properly spelled out, places stringent requirements on c that considerably narrow down the field of potential confirmation measures. In fact, I will show that only the log-likelihood measure meets the requirements. My argument does not, however, establish that the log-likelihood measure is an interval measure, nor that it is the *true* measure of confirmation; the argument only shows that the log-likelihood is the *only* candidate interval measure. This leaves it open that there is no adequate confirmation measure that is an interval measure.

I start by laying out my background assumptions in Section 2. In Section 3, I make the requirements on c more precise. In Section 4, I show how these requirements entail that c is the log-likelihood measure. In Section 5, I discuss the implications of the argument and consider a couple of objections.

2 Background Assumptions

According to a criterion of confirmation universally agreed upon among Bayesians, E *confirms* H just in case $Pr(H|E) > Pr(H)$.[1] Although this criterion suffices to

[1] *Disconfirmation* happens when the inequality sign is reversed, and when there is an equality sign we have *confirmation neutrality*.

answer the binary question of whether or not E confirms H, it does not answer the quantitative question of whether E confirms H to a high degree, nor does it answer the comparative question of which of two hypotheses is confirmed more by E.[2] In order to answer either of the preceding types of question, one needs a confirmation measure that quantifies the degree to which E confirms (or disconfirms) H. The following is a small sample of the measures that have been offered in the literature:

The plain ratio measure, $r(H, E) = \frac{Pr(H|E)}{Pr(H)}$

The log-ratio measure, $lr(H, E) = \log r(H, E)$

The difference measure, $d(H, E) = Pr(H|E) - Pr(H)$

The log-likelihood measure, $l(H, E) = \log(\frac{Pr(E|H)}{Pr(E|\neg H)})$

The alternative difference measure, $s(H, E) = Pr(H|E) - Pr(H|\neg E)$[3]

Since Bayesians analyze confirmation in terms of probability, and since the probability distribution over the algebra generated by H and E is determined by $Pr(H|E)$, $Pr(H)$, and $Pr(E)$, it has become standard to assume that any confirmation measure can be expressed as a function of $Pr(H|E)$, $Pr(H)$, and $Pr(E)$. The preceding assumption is essentially the requirement that Crupi et al. (2013) call "formality." A strong case can however be made for not allowing our measure of confirmation to depend on $Pr(E)$. As Atkinson (2009) points out, if we let $c(H, E)$ be a function of $Pr(E)$, then $c(H, E)$ can change even if we add to E a piece of irrelevant "evidence" E' that is probabilistically independent of H and E, and of their conjunction. To see this, suppose that $c(H, E) = f(Pr(H), Pr(H|E), Pr(E))$. Let E' be any proposition whatsoever that is independent of H, E, and $H\&E$.[4] Then $c(H, E\&E') = f(Pr(H), Pr(H|E\&E'), Pr(E\&E')) = f(Pr(H), Pr(H|E), Pr(E)Pr(E'))$. If f depends on the third argument, we can find some probability function Pr such that $f(Pr(H), Pr(H|E), Pr(E)Pr(E')) \neq f(Pr(H), Pr(H|E), Pr(E))$, and thus such

[2]Carnap (1962) was the first philosopher to draw the distinction between these three questions.

[3]This measure is also sometimes called the "Joyce-Christensen measure," after Joyce (1999) and Christensen (1999).

[4]I am of course assuming here that H and E are fixed.

that $c(H, E\&E') \neq c(H, E)$. However, this is clearly counterintuitive, since E' is probabilistically independent of H and E and therefore should not have any impact on the confirmation of H. So we conclude that f should not depend on $Pr(E)$.

Since I find the preceding argument convincing, I will assume that the confirmation measure we are looking for is of the following form: $c(H, E) = f(Pr(H), Pr(H|E))$. Since there is no *a priori* restriction on what credences an agent may have except that these credences must lie somewhere in the interval $[0, 1]$, I will assume that f is defined on all of $[0, 1] * [0, 1]$. Note that, as Huber (2008) points out, this is not the same as assuming that any particular probability distribution $Pr(*)$ is continuous.

The preceding two assumptions are summed up in the following requirement:

Strong Formality (SF). *Any confirmation measure is of the following form:* $c(H, E) = f(Pr(H), Pr(H|E))$*, where* f *is a function defined on all of* $[0, 1] * [0, 1]$.

It should be noted that (SF) excludes some of the confirmation measures that have been offered in the literature.[5] I briefly address lingering objections to (SF) in Section 5.

Finally, I will also adopt the following convention:

Confirmation Convention (CC).

$$c(H, E) : \begin{cases} > 0 & \text{if } Pr(H|E) > Pr(H), \\ = 0 & \text{if } Pr(H|E) = Pr(H), \\ < 0 & \text{if } Pr(H|E) < Pr(H). \end{cases}$$

(CC) is sometimes taken to be part of the definition of what a confirmation measure is (e.g. by Fitelson (2001)). Although I think it is a mistake to think of (CC) in this way, I will adopt (CC) in this paper for convenience. (CC) has the role of setting 0 as the number that signifies confirmation neutrality.

[5] In particular, the alternative difference measure.

3 The Main Requirement on c

Suppose we witness a coin being flipped 10 times, and our task is to assign a credence to the proposition that the coin comes up heads on the 11th flip. If we do not in advance know anything about the coin's bias, it is reasonable to guess that the coin will come up heads with probability $H/10$ on the 11th flip, where H is the number of times the coin comes up heads in the 10 initial flips.[6] In making this guess, we are setting our credence in the coin landing heads equal to the observed frequency of heads. This move is reasonable since the law of large numbers guarantees that the observed frequency of heads converges in probability to the coin's actual bias. The observed frequency of heads does not necessarily equal the coin's bias after just 10 flips, however. In fact, statistics tells us that the confidence interval around the observed frequency can be approximated by $\hat{p} \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$, where $\hat{p}$ is the observed frequency, n is the sample size (in this case, 10 coin flips), and z is determined by our desired confidence level.

For example, suppose we witness 4 heads in 10 coin flips and we set our confidence level to 95%. In that case, $z = 1.96$, $\hat{p} = 0.4$, and the calculated confidence interval is approximately $[0.1, 0.7]$. Clearly, the confidence interval in this case is rather large. Given our evidence, we can do no better than to estimate the coin's bias as 0.4. However, we also need to realize that if the 10 flips were repeated, we would probably end up with a slightly different value for $\hat{p}$: we should acknowledge that credences are bound to vary with our varying evidence.

The above example illustrates one way that variability can sneak into our credences: if our credence is calibrated to frequency data, then our credence inherits the variability intrinsic to the frequency data. However, even if we set our credence by other means than frequency data, we must admit that rational credences are intrinsically somewhat variable. For example, if the sky looks ominous and I guess that there is a 75% chance that it is going to rain (or perhaps my betting behavior reveals that this is my credence that it is going to rain), I must concede that another agent whose credence (or revealed credence) is 74% or 76% is just as rational as I

[6] This assumes $0 < H < 10$.

am: I do not have either the evidence nor the expertise to discriminate between these credences. And even if I do have good evidence as well as expertise, I must admit that I am almost never in a position where I have *all* the evidence, and had I been provided with somewhat different evidence, I would have ended up with a somewhat different credence.

The fact that our credences are variable is a fact of life that any rational agent must face squarely. It is not hard to see that this fact also affects Bayesian confirmation theory. Bayesian confirmation measures are defined in terms of credences, and are therefore infected by the variability inherent in credences. If Bayesian confirmation measures are necessarily affected by variable credences, I contend that we should want a confirmation measure that is affected by such variability in a systematic and predictable way. We should want this even if we only care about the ordinal properties of confirmation measures. Suppose, for instance, that our confirmation measure is very sensitive to minor variations in the prior or the posterior. In that case, if we find out that $c(H, E) > c(H', E')$, we cannot necessarily be confident that H truly is better confirmed by E than H' is by E' because a small variation in our credence in H or H' might well flip the inequality sign so that we instead have $c(H, E) < c(H', E')$. In order to be confident that $c(H, E)$ really is better confirmed than $c(H', E')$, we need to be assured that the inequality sign is stable. Now, we can be assured that the inequality is stable as long as $c(H, E) - c(H', E')$ is of "significant size." But in order for us to be able to determine that $c(H, E) - c(H', E')$ is "of significant size," we need to be able to draw meaningful and robust conclusions from this difference.

Thus, even if we are primarily interested in the ordinal ranking of evidence-hypothesis pairs provided by c, we still want to be able to draw conclusions from the difference $c(H, E) - c(H', E')$. However, if c is very sensitive to small variations in the priors or posteriors of H and H', then the quantity $c(H, E) - c(H', E')$ is unstable: it could easily have been different, since our priors or posteriors could easily have been slightly different (for instance if we calibrated our priors to frequency data). We are therefore only justified in interpreting the difference $c(H, E) - c(H', E')$ if c is relatively *insensitive* to small variations in the priors and posteriors.

Suppose, moreover, that slight variations in small priors (or posteriors) have a larger effect on c's output than do slight variations in larger priors. Then we cannot compare the quantity $c(H, E) - c(H', E)$ to the quantity $c(H'', E) - c(H', E)$ unless our prior credences in H'' and H are approximately the same. In order for us to be able to compare $c(H, E) - c(H', E)$ to $c(H'', E) - c(H', E)$ in cases where our prior credences in H'' and H are very different, we need c to be *uniformly* insensitive to small variations in the prior (and the posterior). We can sum up the preceding two remarks as follows:

Main Requirement (MR). *We are justified in interpreting and drawing conclusions from the quantity $c(H, E) - c(H', E')$ only if c is uniformly insensitive to small variations in $Pr(H)$ and $Pr(H|E)$.*

As it stands, (MR) is vague. What counts as a small variation in a credence? Moreover, what does it mean, concretely, for c to be *uniformly insensitive* to such variations? To get a better handle on these questions, let us formalize the important quantities that occur in (MR). Following (SF), we are assuming that c is of the form $c(H, E) = f(Pr(H), Pr(H|E))$. For simplicity, let us put $Pr(H) = x$ and $Pr(H|E) = y$, so that $c = f(x, y)$. According to (MR), we require that f be uniformly insensitive to small variations in x and y. I will use $v(p, \epsilon)$ to capture the notion of a small variation in the probability p, where ϵ is a parameter denoting the size of the variation. Moreover, I will use $\Delta^x_\epsilon c(x, y)$ to denote the variation in c that results from a variation of size ϵ about x. That is to say,

$$\Delta^x_\epsilon c(x, y) = f(v(x, \epsilon), y) - f(x, y) \tag{3.1}$$

Similarly, I will use $\Delta^y_\epsilon c(x, y)$ to denote the variation in c that results from a variation of size ϵ about y. Thus,

$$\Delta^y_\epsilon c(x, y) = f(x, v(y, \epsilon)) - f(x, y) \tag{3.2}$$

The next step is to get a better grip on (MR) by investigating the terms that occur in (3.1) and (3.2). In sections 3.1 through 3.3, that is what I do.

3.1 What is uniform insensitivity?

First, the demand that c be *uniformly* insensitive to variations in the prior and the posterior now has an easy formal counterpart: it is simply the demand that for different values x_1,x_2,y_1, and y_2 of x and y, we have $\Delta_\epsilon^{x_1} c(x_1, y_1) = \Delta_\epsilon^{x_2} c(x_2, y_2) = \Delta_\epsilon^{x_2} c(x_2, y_1) = etc.$ and $\Delta_\epsilon^{y_1} c(x_1, y_1) = \Delta_\epsilon^{y_2} c(x_2, y_2) = \Delta_\epsilon^{y_2} c(x_1, y_2) = etc.$ Thus, across different values of x and y, a small variation in c will mean the same thing. More importantly, this means we can consider $\Delta_\epsilon^x c(x, y)$ as purely a function of ϵ, and likewise for $\Delta_\epsilon^y c(x, y)$. From now on, I will therefore write:

$$g(\epsilon) := \Delta_\epsilon^x c(x, y) \tag{3.3}$$

$$h(\epsilon) := \Delta_\epsilon^y c(x, y) \tag{3.4}$$

In order to figure out what the requirement that c be insensitive to *small* variations amounts to, we need to figure out how to quantify variations in credences. It is to this question that I now turn.

3.2 What is a small variation in a credence?

Given a credence x, what counts as a small variation in x? This question turns out to have a more subtle answer than one might expect. Using the notation from equations 3.1 and 3.2, what we are looking for is the form of the function $v(x, \epsilon)$. Perhaps the most natural functional form to consider is the following one: $v(x, \epsilon) = x + \epsilon$. On this model, a small (positive) variation in the probability x is modeled as the addition of a (small) number to x. However, if we consider specific examples, we see that this model is too crude. For example, supposing that $x = 0.5$, we might consider 0.05 a small variation relative to x. But if we consider $x = 0.00001$ instead, then 0.05 is no longer small relative to x; instead it is now several orders of magnitude bigger.

The above example shows that the additive model cannot be right. An easy fix is to scale the size of the variation with the size of x. In other words, we might suggest the following form for v: $v(x, \epsilon) = x + x\epsilon$. This adjustment solves the problem mentioned in the previous paragraph. According to the new v, a variation of size

0.025 about 0.5 is "equal" to a variation of 0.0000005 about 0.00001. In contrast to the previous additive model, $v(x, \epsilon) = x + x\epsilon$ is a "multiplicative" model of variability, as we can see by instead writing it in the following form: $v(x, \epsilon) = x(1 + \epsilon)$

However, the multiplicative model, though much better than the additive model, is still insufficient. One problem is purely mathematical. Since $v(x, \epsilon)$ is supposed to correspond to a small positive shift in probability, we should require that $0 \leq v(x, \epsilon) \leq 1$, for all values of x and ϵ. However, $x + x\epsilon$ can easily be larger than 1, for example if $x = 0.9$ and $\epsilon = 0.2$.[7] The other problem is that $v(x, \epsilon)$ treats values of x close to 0 very differently from values of x close to 1. For instance, a variation where $\epsilon = 0.1$ will be scaled to 0.001 when $x = 0.01$. But when $x = 0.99$, the same ϵ will be scaled to just 0.099. This is very problematic, since for every hypothesis H in which we have a credence of 0.99, there corresponds a hypothesis in which we have a credence of 0.01, namely $\neg H$. But a small variation in our credence in H is necessarily also a small variation in our credence in $\neg H$, simply because $Pr(\neg H) = 1 - Pr(H)$: H and $\neg H$ should therefore be treated symmetrically by v. There is an easy fix to both of the preceding problems: if we scale ϵ by $x(1 - x)$ instead, then first of all we have $0 \leq x + \epsilon x(1 - x) \leq 1$, and thus $0 \leq v(x, \epsilon) \leq 1$. Second of all, H and $\neg H$ are now treated symmetrically. From the preceding considerations, we therefore end up with the following as our functional form for v: $v(x, \epsilon) = x + x(1 - x)\epsilon$.

There is a completely different argument by which we can arrive at the same functional form for v. As I mentioned in the example at the beginning of section 3, credences are sometimes calibrated to frequency data. This is for example usually the case if H is a medical hypothesis. Suppose H represents the hypothesis that a person P has disease X, for instance. The rational prior credence in H (before a medical examination has taken place) is then the frequency of observed cases of X in the population from which P is drawn. The frequency of observed cases of X can be modeled as the outcome of a binomial process having mean $Pr(H)$ and variance $Pr(H)(1 - Pr(H))$. Suppose we observe the frequency $fr(\hat{H})$. Then the estimated variance is $Var(H) \approx fr(\hat{H})(1 - fr(\hat{H}))$. The variance is maximal at $fr(\hat{H}) = 0.5$ and decreases as $fr(\hat{H})$ moves closer to 0 or to 1. Arguably, it makes a

[7] This is also a problem for the additive model

lot of sense in this case for the variability in one's credence to vary with the variance of the frequency data. But that is exactly what $v(x, \epsilon) = x + x(1 - x)\epsilon$ does: it scales credence variability by data variance. Thus, according to v, a variation of size $Var(X)\epsilon$ about credence X is equal to a variation of size $Var(Y)\epsilon$ about credence Y.

From all the preceding considerations, I conclude that the following is the best functional form for v:

$$v(x, \epsilon) = x + x(1 - x)\epsilon \tag{3.5}$$

3.3 Uniform insensitivity to small variations in the prior and posterior

The next step is to understand what insensitivity amounts to. To say that c is *insensitive* to small variations in the prior or posterior is to say that such variations have a small effect on confirmation: the most natural way to formalize this requirement is in terms of continuity. Since $g(\epsilon)$ represents the change in confirmation resulting from a change (by ϵ) in probability, a natural continuity requirement for c would be that g and h should be continuous at 0.

However, continuity is too weak a requirement. Even if a function is continuous, it is still possible for it to be very sensitive to small variations. For instance, the function $f(x) = 1000000^x$ is continuous (everywhere), but is at the same time very sensitive to small perturbations of x. Sensitivity to perturbations is most naturally measured by looking at how the derivative behaves. Minimally, we should therefore require that g and h be differentiable at 0. The next natural requirement would be to demand that the derivative of both g and h be *bounded* by some "small" number. Of course, pursuing such a requirement would require a discussion of what is to count as a "small" number in this context. Since I do not actually need a requirement of this sort in my argument in the next section, I will not pursue a discussion of these issues here. The only upshot from this section is therefore that g and h should be differentiable at 0.

4 The Main Result

Let me summarize where we are. Our desire to be able to draw conclusions from differences in confirmation, i.e. from expressions of the form $c(H, E) - c(H', E')$, led us to the requirement that c be *uniformly insensitive to small variations* in $Pr(H)$ and $Pr(H|E)$. In sections 3.1 through 3.3, I made the various components of this requirement more precise. Putting all these components together, we have the following:

Formal Version of the Main Requirement (MR) 4.1. *We are justified in drawing conclusions from the difference $c(H, E) - c(H', E')$ only if the following conditions are all met:*

1. $f(v(x, \epsilon), y) - f(x, y) = g(\epsilon)$, *where:*

2. $g(\epsilon)$ *does not depend on either x or y*

3. $g(\epsilon)$ *is differentiable at 0*

4. $v(x, \epsilon) = x + x(1 - x)\epsilon$

5. $f(x, v(y, \epsilon) - f(x, y) = h(\epsilon)$, *where:*

6. $h(\epsilon)$ *does not depend on either x or y*

7. $h(\epsilon)$ *is differentiable at 0*

Note that (5) - (7) are just (1) - (3) except that they hold for h instead of for g. Note also that (MR) is essentially *epistemic*. It says that "we" (i.e. agents

interested in confirmation) are only justified in drawing conclusions (of any kind) from $c(H, E) - c(H', E')$ if certain formal conditions are met. These conditions ensure that $c(H, E)$ behaves reasonably well. Together with (SF) and (CC), the conditions in (MR) entail the log-likelihood measure, as I show next.

Main Result 4.1. *If (MR) is true, (SF) is assumed, and (CC) is adopted as a convention, then*

$$c(H, E) = \log \frac{Pr(E|H)}{Pr(E|\neg H)}$$

Where the identity is unique up to positive linear transformations with constant term 0.

Proof. Starting with (1) from (MR), we have,

$$f(v(x, \epsilon), y) - f(x, y) = g(\epsilon) \tag{4.1}$$

If we divide each side by $x(1 - x)\epsilon$, we get:

$$\frac{f(v(x, \epsilon), y) - f(x, y)}{x(1 - x)\epsilon} = \frac{g(\epsilon)}{x(1 - x)\epsilon} \tag{4.2}$$

Next, we let $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} \frac{f(v(x, \epsilon), y) - f(x, y)}{x(1 - x)\epsilon} = \lim_{\epsilon \to 0} \frac{g(\epsilon)}{x(1 - x)\epsilon} \tag{4.3}$$

Since g is differentiable at 0 (from part (3) of (MR)), the right hand side of the above equation is just $\frac{1}{x(1-x)} g'(0)$. Since the limit exists on the right hand side of the equation, it must exit on the left side as well. But the left side is just $\frac{\partial}{\partial x} f(x, y)$. We therefore have,

$$\frac{\partial}{\partial x} f(x, y) = \frac{1}{x(1 - x)} g'(0) \tag{4.4}$$

Next, we take the antiderivative of each side of (4.4) with respect to x. Since g and hence $g'(0)$ does not depend on x (from part (2) of (MR)), we have:

$$f(x,y) = g'(0)(\log x - \log(1-x)) + C \tag{4.5}$$

Here, C is a number that depends on y but not on x. If we perform the above calculations again starting instead with $f(x, v(y,\epsilon)) - f(x,y) = h(\epsilon)$, we find that:

$$C = h'(0)(\log y - \log(1-y)) + K \tag{4.6}$$

Here, K is just a constant, i.e. it depends on neither x nor y. We therefore have:

$$f(x,y) = g'(0)(\log x - \log(1-x)) + h'(0)(\log y - \log(1-y)) + K \tag{4.7}$$

Now set $x = y = 0.5$. The second part of (CC) then entails that $K = 0$. Next, set $x = y$. Then (CC) entails:

$$g'(0)(\log x - \log(1-x)) + h'(0)(\log x - \log(1-x) = 0 \tag{4.8}$$

This in turn entails that $g'(0) = -h'(0)$. Thus we have,

$$f(x,y) = -h'(0)(\log x - \log(1-x)) + h'(0)(\log y - \log(1-y)) \tag{4.9}$$

$$= h'(0) \log \frac{y}{1-y} * \frac{1-x}{x} \tag{4.10}$$

Remembering that $x = Pr(H)$ and $y = Pr(H|E)$, (4.9)-(4.10) together with (SF) entail:

$$c(H,E) = f(Pr(H), Pr(H|E)) \tag{4.11}$$

$$= h'(0) \log \frac{Pr(H|E)}{1 - Pr(H|E)} * \frac{1 - Pr(H)}{Pr(H)} \tag{4.12}$$

$$= h'(0) \log \frac{Pr(H|E)}{Pr(H)} * \frac{Pr(\neg H)}{Pr(\neg H|E)} \tag{4.13}$$

$$= h'(0) \log \frac{Pr(H|E) * Pr(E)}{Pr(H)} * \frac{Pr(\neg H)}{Pr(\neg H|E) * Pr(E)} \tag{4.14}$$

$$= h'(0) \log \frac{Pr(E|H)}{Pr(E|\neg H)} \tag{4.15}$$

Finally, (CC) entails that $h'(0)$ must be a positive number. Thus $c(H,E) = l$, up to positive linear transformations with constant term 0.

□

5 Discussion and Objections

In the previous section, I showed that (MR), (SF), and (CC) jointly entail the log-likelihood confirmation measure, l. The proof entails l up to strictly positive linear transformations with constant term 0. That is to say, if $\log \frac{Pr(E|H)}{Pr(E|\neg H)}$ is a legitimate confirmation measure, then so is $a * \log \frac{Pr(E|H)}{Pr(E|\neg H)}$, for $a > 0$; the argument does not establish that any particular logarithmic base is better than another. In Stevens (1946)'s terminology, our measure is apparently *ratio*, meaning that we are justified in interpreting both intervals and ratios between outputs of the measure. Analogously, mass is also a ratio measure since it makes sense to say both that the difference between 2kg and 4kg is the same as the difference between 4kg and 6kg, and that 4kg is twice as big as 2kg.

It therefore appears that my conclusion is stronger than what I set out to estab-

lish: in the introduction, I said that the goal was to find a confirmation measure that can be interpreted as an interval measure. But the proof in the previous section apparently establishes that l is a ratio measure. However, contrary to the appearances, I think it is illegitimate to interpret l as a ratio measure. The difference between interval measures and ratio measures is that ratio measures have a non-arbitrary 0. But in our case, it is (CC) that establishes our measure's 0, and (CC) is (as the name suggests) just a convention. We could just as easily have chosen a convention such that 1 meant confirmation neutrality. Therefore, the 0 is arguably arbitrary, and it is not legitimate to interpret our measure as anything more than an interval measure.

The second thing to notice about my argument is that it does not actually establish that the log-likelihood measure is the *true* confirmation measure. This is because (MR) merely gives necessary conditions, and no sufficient ones. Thus, what my argument shows is really a conditional statement: *if* there is any interval confirmation measure, then that measure is l. The preceding conditional is, of course, equivalent to the following disjunction: *either* there is no interval confirmation measure *or* the only interval confirmation measure is l.

The third and final observation I will make about the argument is that it clearly depends very much on the choice of v. In Section 3.2 I considered and rejected two other measures of variability: the additive measure, $v(x, \epsilon) = x + \epsilon$; and the multiplicative measure, $v(x, \epsilon) = x + x\epsilon$. It is natural to ask what confirmation measures we end up if we instead use these alternative measures of credence variability. The answer, although I will not show this here, is that the additive measure yields the difference confirmation measure, d, whereas the multiplicative measure yields the log-ratio confirmation measure, lr. We can therefore see that d and lr "embody" defective measures of credence variability: arguably, that is a strike against these measures.

Next, I will consider a couple of objections to the argument. First, my argument is obviously only sound if the assumptions in (MR) are correct. However, the assumptions in (MR) might remind the reader of assumptions made in Good (1960, 1984) and Milne (1996). These assumptions have been criticized by Fitelson

(2001, 2006) as being "strong and implausible" (2001, 28-29n43) and for having "no intuitive connection to material desiderata for inductive logic" (2006, 7n13).

Why does my argument escape Fitelson's criticisms? How is my argument different from the arguments made by Good and Milne? The answer is that, whereas Good and Milne are not interested in the *interval* properties of their confirmation measures, and the various mathematical assumptions they make therefore seem unmotivated, all the properties listed in (MR) arise naturally out of our wish to have a confirmation measure that is at least an interval measure.

Finally, one may object to some of the other background assumptions I make in Section 1. In particular, *Strong Formality* may be accused of being too strong since it excludes the alternative difference measure right off the bat. My reply to this objection is as follows: the argument in Section 4 can be carried out without *Strong Formality*, but the resulting analysis does not yield the alternative difference measure, nor any other recognizable confirmation measure. Thus, even if one rejects (SF), one cannot use the type of argument I have given in this paper to argue for the alternative difference measure or other standard measures that depend non-trivially on $Pr(E)$.[8]

6 Conclusion

I have argued that there is a set of conditions that any confirmation measure must meet in order to justifiably be interpreted as an interval measure. Furthermore, I have shown that these necessary conditions, together with an additional plausible assumption and a widely accepted convention, jointly entail the log-likelihood measure. My argument does not show that l is an interval measure, but it does show that it is the only measure that stands the chance of being one. Nor does the argument in this paper show that l is the "true" confirmation measure. However, to the extent that we care about our measure's being an interval measure, we should regard the conclusion in this paper as favoring l as our preferred measure.

[8]Such as Carnap's measure, $c(H, E) = Pr(H\&E) - Pr(H)P(E)$.

References

Atkinson, D. (2009). Confirmation and justification. A commentary on Shogenji's measure. *Synthese*, 184(1):49–61.

Carnap, R. (1962). *Logical Foundations of Probability*. Chicago: University of Chicago Press, second edition.

Christensen, D. (1999). Measuring confirmation. *Journal of Philosophy*, XCVI:437–61.

Crupi, V., Chater, N., and Tentori, K. (2013). New axioms for probability and likelihood ratio measures. *British Journal for the Philosophy of Science*, 64:189–204.

Fitelson, B. (2001). *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin – Madison.

Fitelson, B. (2006). Logical foundations of evidential support. *Philosophy of Science*, 73:500–12.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society: Series B*, 22:319–31.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation*, 19:294–299.

Huber, F. (2008). Milne's argument for the log-ratio measure. *Philosophy of Science*, 75:413–20.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Milne, P. (1996). Log[P(hleb)/P(hlb)] is the one true measure of confirmation. *Philosophy of Science*, 63:21–6.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, 103(2684):577–80.

DRAFT—PLEASE DO NOT CITE. COMMENTS ARE WELCOME

Paper to be presented at the PSA 2014, Chicago, 6-8 November 2014

Symposium: "How Adequate Are Causal Graphs and Bayesian Networks?"

***

On the Incompatibility of Dynamical Biological Mechanisms and Causal Graph Theory

Marcel Weber
Department of Philosophy
University of Geneva
marcel.weber@unige.ch

Abstract

I examine the adequacy of the causal graph-structural equations approach to causation for modeling biological mechanisms. I focus in particular on mechanisms with complex dynamics such as the PER biological clock mechanism in *Drosophila*. I show that a quantitative model of this mechanism that uses coupled differential equations – the well-known Goldbeter model – cannot be adequately represented in the standard (interventionist) causal graph framework, even though this framework does permit causal cycles. The reason is that the model contains dynamical information about the mechanism that concerns causal properties but that does not correspond to variables that could be subject to independent interventions. Thus, a representation of the mechanisms as a causal structural model necessarily suppresses causally relevant information.

1. Introduction

Recent decades have seen the advent of elaborate formal techniques for causal modeling (Spirtes, Glymour, and Scheines 2000; Pearl 2000). These techniques, which essentially link causality to manipulability, have been instrumental in taking philosophical debate about causation as well as about scientific explanation to a new level (e.g., Woodward 2003, 2011; Woodward and Hitchcock 2003; Hitchcock and Woodward 2003; McKay Illari, Russo and Williamson 2011). Furthermore, this formal approach to causality has been productively applied in order to analyze causation in specific scientific disciplines. Originally developed mainly in the context of econometrics, it was recently also applied to various other sciences, e.g., neuroscience (Craver 2007, Weber 2008), genetics (Waters 2007; Woodward 2010), evolutionary theory (Otsuka forthcoming), psychiatry (Woodward 2008), or public health policy (Russo 2012), to name just a few.

The basic tools of this approach are the formally definable concepts of directed acyclic graph (DAG), Bayesian network, and structural equation. In the standard approach, the causal interpretation of these formal concepts is provided by means of the concept of idealized intervention. The result are causal models that contain information about counterfactual dependencies between a set of variables as well as, in some cases, probability distributions defined over these variables.

While the fruitfulness of this approach to causal modeling in scientific practice as well as for philosophical analysis is beyond doubt, there have not been many attempts to explore its limits in adequately representing causal systems. There has, of course, been

quite some debate concerning the question of whether a certain conception of *mechanism* is adequate for explaining biological phenomena (e.g., Bechtel 2005, 2013; Bechtel and Abrahamsen 2010; Braillard 2010; Kuhlmann 2011; Waskan, 2011; Weber 2012; Dupré 2013; Woodward 2013). This debate focused on the issue of whether the standard conceptions of mechanism can account for biological processes that feature complex dynamical behavior and/or spatial structures. However, none of this work has directly challenged the underlying interventionist theory of causation itself. In fact, there is a whole range of more recent studies that attempt to show that Bayesian networks are actually adequate for modeling complex biological mechanisms (Casini et al. 2011; Clarke, Leuridan and Williamson 2014; Gebharter 2014; Gebharter and Kaiser 2014; Gebharter and Schurz, this symposium; Casini and Williamson, this symposium).

In part, this problem turns on the question of how narrowly the term "mechanism" should be understood (Woodward 2013). In this paper, I will not be concerned with this issue. Rather, I want to examine to what extent the contemporary interventionist approach to causality is apt for representing the causal properties of a certain kind of mechanism in the first place.

A critical issue will be the extent in which causal models that basically contain causal difference-making information can account for the dynamics and for spatial features of mechanisms, as such features are absolutely crucial for the explanatory force of many mechanisms, in biology and elsewhere. Woodward (2013) has argued that spatio-temporal information can always be integrated with the causal difference-making information

contained in causal models. While this may be true in some sense, I will show that it glosses over a basic problem pertaining to the dynamics of certain kinds of causal system.

I will closely examine an example from biology that involves a mechanistic model consisting of a system of coupled differential equations with complex dynamics. This model describes the operation of a biological clock. I will assume without further argument that this model captures the essential causal properties of the biological clock mechanism, at least with respect to certain explanatory goals.[1] Then, I will show that formal causal models fail to correctly represent these causal properties. Specifically, I will argue that such a model will not be able to treat time derivatives as causally relevant variables.

I shall proceed as follows. In Section 2, I shall briefly review the core notions used in the causal modeling literature, in particular the notions of causal graph, structural equations, and ideal intervention. In Section 3, I analyze a dynamical model of a biological clock mechanism and show that it has no adequate causal graph representation. In Section 4, I consider some attempts from the current causal modeling literature to represent differential equations in structural causal models. I show that the results coming from these

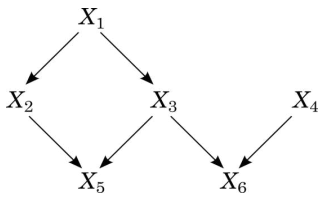
[1] I am not assuming that there is just one correct way of representing a causal system. Thus, I accept the pluralist thesis according to which there is always a variety of different perspectives on the world none of which succeeds in providing a complete picture (Kellert, Longino and Waters 2006; Dupré 2013). In fact, I suggest that my arguments presented here could be used for actually defending such a strong form of scientific pluralism, but this would go beyond the scope of this paper.

attempts actually support my thesis. Section 5 summarizes and integrates my conclusions with regard to the limitations of causal modeling.

2. Causal Modeling

The formal concepts used in causal modeling include directed acyclic graphs (DAGs), structural equations and Bayesian Networks. In this paper, I shall focus on DAGs and structural equations and leave Bayesian networks aside, but it should be noted that any problem concerning DAGs will also affect causally interpreted structural equations as well as Bayesian networks because the latter two kinds of causal models contain DAGs.[2]

A DAG is an ordered pair $\langle V, E\rangle$, where V is the set of variables and E is a set of directed edges.



A DAG becomes a causal graph as soon as its edges are interpreted causally, about which I will say a little more below.

Causal dependencies can also be represented by using so-called *structural models* (Pearl 2000). Such a model consists of an ordered triple $\langle U, V, Q\rangle$ where U is a set of exogenous variables, V a set of endogenous variables, and Q a set of *structural equations*. The structural equations give the value of each endogenous variable as a function of the

[2] I wish to thank Lorenzo Casini for pointing this out to me.

values of other variables in *U* and *V*. The variables may also be interpreted as nodes that are connected by causal arrows. But in contrast to pure causal graphs, the structural equations also provide quantitative information as to how much some dependent variables change per unit change of the independent variable.

Pearl (2000, p. 160) gives the following "operational" definition of a structural equation:

> *An equation* $y = \beta x + \varepsilon$ *is said to be structural if it is to be interpreted as follows: In an ideal experiment where we control* X *to* x *and any other set* Z *of variables (not containing* X *or* Y*) to z, the value* y *of* Y *is given by* $\beta x + \varepsilon$*, where ε is not a function of the settings* x *and* z*.*

According to this definition, it is obvious that structural equations *sensu* Pearl are *linear* equations in the sense of not containing derivatives of the variables. As we shall see, this feature constitutes a major limitation when it comes to modeling systems with complex dynamics.

Pearl's definition of a structural equation contains the idea of an "ideal experiment". This notion has been elaborated in great detail by Woodward (2003, 94-99), who defines it in terms of the notion of ideal intervention. On this account, an (ideal) intervention on some variable *X* with respect to some variable *Y* changes *Y* by changing *X* without changing any other variable that is a cause of *Y*.

In a nutshell, these are the basic concepts of causal modeling. Thus, when I speak about a "causal model" in what follows, I mean a model that is expressed by using either causal graphs or structural equations and that uses an interventionist criterion for

interpreting the graphs and equations causally. The goal of this paper (as well as the paper by Kaiser, this symposium) is to show that these concepts fail to fully account for certain *causal* explanations in biology.

In the following section, I show what problems are created for causal models by complex dynamical information. Kaiser (this symposium) does the same for spatially complex mechanisms. Thus, while Kaiser's paper is about space, this one is about time.

3. It's About Time: Modeling Dynamic Processes

3.1. Classic Examples of Dynamic Models in Biology

There is an important class of biological models that try to account for complex series of events in dynamical terms. A classic example is the Hodgkin-Huxley model of the action potential (see, e.g., Weber 2005, 2008). This model (henceforth HH model) shows how changes in membrane conductance generate a temporary membrane depolarization that can spread along an axonal membrane and thus form the basis of information processing by neurons. A more recent example is Goldbeter's (1995) model of the circadian oscillations of the PER protein in *Drosophila*, which is the heart of a biological clock mechanism. There are many more such models, but for the purposes of this paper we shall concentrate on these two.

3.2. Bechtel and Abrahamsen on Dynamic Mechanistic Explanation

In a recent series of papers, Bill Bechtel and Adele Abrahamsen have provided a very illuminating account of models and mechanisms in circadian clock research, including

Goldbeter's model and the PER mechanism (Bechtel and Abrahamsen 2010, Bechtel 2013). Their account will prove to be useful for our analysis, which is why it will be briefly reviewed here. We take the gist of their account to be that circadian clock models provide what they call *dynamic mechanistic explanations*. According to Bechtel and Abrahamsen, such explanations differ from other kinds of mechanistic explanations in providing *quantitative* information about the behavior of the systems in question. Dynamic mechanistic explanations (may) contain *sequential mechanistic models* that describe a series of events in purely *qualitative* terms. Figure 1 shows such a sequential mechanistic model.

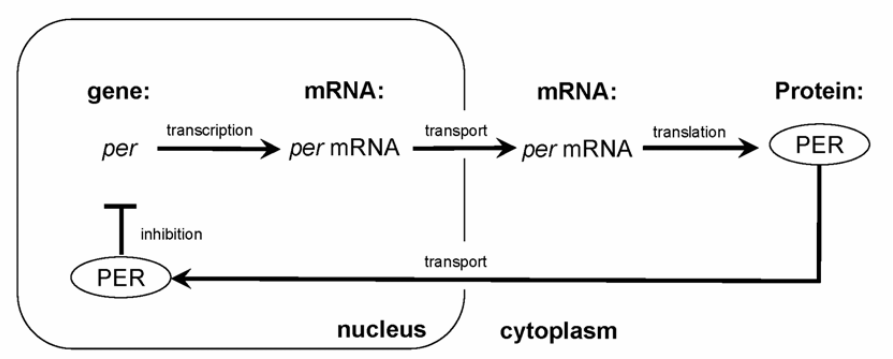


Figure 1. The sequential mechanism of the *Drosophila* circadian clock gene *period*. After Hardin et al. (1990).

An interesting feature of this sequential model according to Bechtel and Abrahamsen is the fact that it is possible to *mentally rehearse* the individual steps as well as their temporal arrangement.

But the most important claim made by Bechtel and Abrahamsen for our purposes is the following: The qualitative sequential model as shown in Figure 1 is *incomplete*. For what the model must show is that the circadian system is capable of generating *stable* oscillations. This is where the dynamical, quantitative model constructed by (Goldbeter 1995) comes in. The model describes the change in cytoplasmic concentrations of PER mRNA (M) as well as the different phosphorylation states of cytoplasmic (P_0, P_1, P_2) as well as nuclear (P_N) PER protein with the help of differential equations. The model uses standard Michaelis-Menten enzyme kinetics where the V_i are maximal reaction rates and the K_i the so-called Michaelis constants for the different biochemical reactions involved (the Michaelis constant gives the substrate concentration at which the reaction rate is half the maximal rate).

The structure of the dynamical model can be extracted from Figure 5.

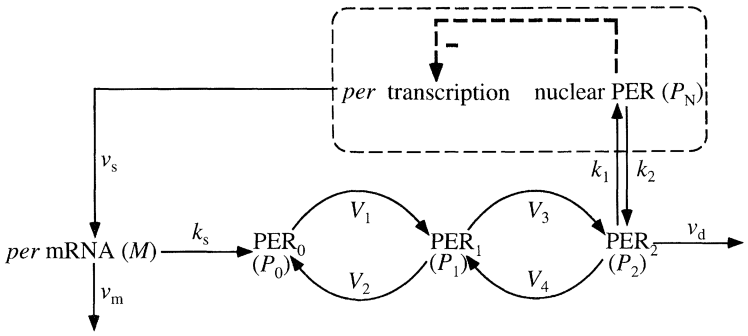


Figure 2. The structure of Goldbeter's dynamical model (after Goldbeter 1995). The concentration of *per* mRNA is represented by M, that of different forms of the PER protein by P_i. P_0 is the unphosphorylated, P_1 the monophosphorylated and P_2 the biphosphorylated form. P_N is for the nuclear PER protein, all the other concentrations are cytosolic. v_s is the

maximal rate of mRNA synthesis, v_m and K_m are the maximum rate and Michaelis constant for the enzymatic degradation reaction of the mRNA. The V_i and K_i give the maximum rates and Michaelis constants for the kinases and phosphatases catalyzing reversible phosphorylation reactions. v_d and K_d are the enzymatic parameters for the degradation reaction of fully phosphorylated PER. k_1 is a rate constant for the transport of PER protein into the cell nucleus, k_2 for the reverse transport. Feedback inhibition of *per* mRNA by nuclear PER is modeled by a Hill equation with a cooperativity of n and a repression threshold constant K_I.

Goldbeter wrote down the reaction rates for the different molecular species as follows:

$$\frac{dM}{dt} = v_s \frac{K_I^n}{K_I^n + P_N^n} - v_m \frac{M}{K_m + M} \quad (1a)$$

$$\frac{dP_0}{dt} = k_s M - V_1 \frac{P_0}{K_1 + P_0} + V_2 \frac{P_1}{K_2 + P_1} \quad (1b)$$

$$\frac{dP_1}{dt} = V_1 \frac{P_0}{K_1 + P_0} - V_2 \frac{P_1}{K_2 + P_1} - V_3 \frac{P_1}{K_3 + P_1} + V_4 \frac{P_2}{K_4 + P_2} \quad (1c)$$

$$\frac{dP_2}{dt} = V_3 \frac{P_1}{K_3 + P_1} - V_4 \frac{P_2}{K_4 + P_2} - k_1 P_2 + k_2 P_N - v_d \frac{P_2}{K_d + P_2} \quad (1d)$$

$$\frac{dP_N}{dt} = k_1 P_2 - k_2 P_N \quad (1e)$$

The total (nonconserved) quantity of PER protein, P_t, is given by:

$$P_t = P_0 + P_1 + P_2 + P_N \quad (2)$$

Using numerical integration techniques, Goldbeter was able to show that for some parameter values there is indeed a limit cycle, in other words, a stable oscillation of the concentrations of mRNA and PER protein.

Bechtel and Abrahamsen stress that without this quantitative model, the sequential model provides no explanation for the *stability* of the circadian behavior. Without introducing quantitative parameters, the sequential model could produce all kinds of behavior, only some of which generate a limit cycle. Thus, the dynamical model must complement the sequential model to obtain the full explanation.

I will argue now that *at best* the sequential model *sensu* Bechtel and Abrahamsen can be represented as a causal model. The dynamical model cannot be so represented, even though it clearly represents a *causal process* (in an idealized and simplified way). Thus, I shall argue that the Goldbeter model is a case of a biological explanation that cannot be accounted for by causal graph models.

3.3 The Sequential Model as a Structural Causal Model

I shall first attempt to represent what Bechtel and Abrahamsen call the sequential model within this causal framework. There is an apparent difficulty in that the sequential model is *cyclical* whereas causal graphs are *acyclical*. However, this problem is not new and solutions have been proposed by several authors (Kistler 2013; Gebharter and Kaiser 2014; Clarke, Leuridan and Williamson 2014). Briefly, one way of doing this is by introducing a time index on some of the nodes of the causal graph structures. When a system comes to the end of a cycle, time has passed. This new state of the system should thus be represented

by a different node, a variable that represents the state of the system at a later time. This way, the cyclical path is broken up and "rolled out" in time and presents no problems for the causal modeler.

However, it should be clear that such a causal graph fails to fully explain the explanandum phenomenon, because essential dynamical information is missing. The graph would merely represent what Bechtel and Abrahamsen refer to as the sequential model. In the next section, I shall examine how the dynamical model could be represented.

3.4 The Dynamical Model as A Structural Causal Model

Could the same strategy that works for the sequential model also be used for representing Goldbeter's dynamical model by using causal graphs? It could be suggested that the causal structure of the model is captured by the following time-indexed causal graph:

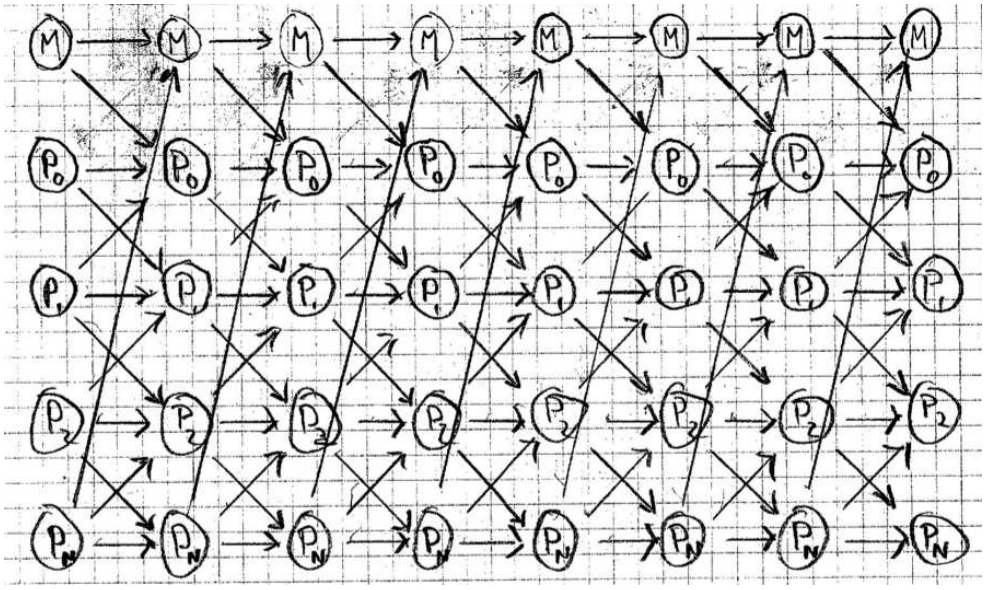


Figure 3. Proposed time-indexed DAG representing the causal dependencies in the Goldbeter model.

It could be argued, perhaps, that this DAG contains all the causal relations posited by the Goldbeter model. A quantitative structural model could also be constructed, for example by writing down rules for updating the values of the salient variables from each discrete time point to the next.

However, it should be clear that such a causal structural model would not be the same as the Goldbeter model. Differential equations with continuous time are

mathematically clearly different from a model with discrete time points.[3] Perhaps there is a discrete-time model that makes approximately the same prediction as Goldbeter's model. In fact, numerical simulations of the equation system use pretty much this strategy. However, the following difficulty arises: In order to really explain the explanandum phenomenon, a model must incorporate temporal information, namely information about how *rates of change* affect the behavior of the system. Goldbeter's differential equations contain precisely such information, and this information is crucial for the model's explanatory force. In fact, I wish to maintain that rates of change are *causally* relevant, because they are important determinants for the behavior of the whole causal process. Thus, I will show now that the Goldbeter model contains causally relevant variables that cannot be represented in the causal graph framework. Furthermore, to the extent that the model is substituted by a discrete-time model that is (approximately) predictively equivalent, the same difficulty arises.

My main argument is that Goldbeter equations do not have the right manipulability properties that are required by structural causal models. I will show, first, that not all causal variables can be subject to *ideal interventions* as required by the causal graph theory. Second, I want to show that the equations do not satisfy the *modularity* requirement that is widely thought to be important in causal models.

First, to see the problem with ideal interventions, consider for example equation (1a) of the Goldbeter model. Suppose we wished to intervene on *M*, the mRNA

[3] It is known that difference equations can have quite surprising properties, see May (1974).

concentration. This obviously cannot be done in a way that leaves the time derivative dM/dt unchanged (if I want to go faster on my bike, thus changing the value of v, I have to accelerate and thus change the value of dv/dt). The same problem occurs for all the other causally relevant variables in the model. Note also that a discrete model faces the exact same difficulty; the only difference is that the rates of change are defined over a time interval instead of a time point. Thus, these equations cannot be subject to the idealized interventions that define causal relations according to causal modelers.

Second, to see the failure of *modularity*, consider for example equations (1b) and (1c). Let us examine what happens when we replace (1b) by the following equation (1b*): $dP_0/dt = p_0$, where p_0 is some real number. This would not only wipe out the r.h.s. of (1b), it would also affect the equations that determine the value of P_0. The reason is, once again, that dP_0/dt and P_0 cannot be manipulated independently of each other. The same problem occurs for the other M- and P-variables. Therefore, the system of equations fails to satisfy the condition of modularity *sensu* Woodward (2003, 48-49, 327-39), which can also be viewed as a kind of manipulability.[4]

What features of the Goldbeter model are responsible for this lack of manipulability, including modularity? It seems to us that the main such feature is the fact

[4] The purpose of the modularity condition is normally to ensure that different equations represent different causal pathways or mechanisms. Perhaps it could be argued that, indeed, some causal mechanisms in the Goldbeter model overlap. For instance, there is a causal cycle between M and P_0 as well as between P_0 and P_1 and these causal cycles share P_0 as a common constituent (cf. Casini and Williamson unpublished).

that some causal variables that occur in the model *affect their own rate of change*, and that for these variables their rate of change is of crucial relevance – indeed *causal* relevance – for the behavior of the whole system. For example, the rate of change of mRNA (*M*) depends on its own concentration. This is due to a *causal* process that is mediated by RNA-degrading enzymes. Furthermore, the concentration of monophosphorylated protein P_1 depends causally on the concentration of unphosphorylated protein P_0, which in turn depends on P_1. Both causal dependencies are mediated by kinases, thus they are *causal* processes.[5]

In Goldbeter's representation of these processes, not only the *values* of the variables at a given time point but also their *rates of change* are *causally relevant*. In other words, it matters not only that a variable X change its value from x_1 to x_2, which is the kind of information that can be encoded in causal graphs. It matters also *how long it takes* for a variable to change by some amount, including an infinitesimally small amount. This rate of change is a *causally relevant property*, but this causal relevance cannot be represented as a causal dependence in the causal framework because the rate of change cannot be

[5] An anonymous referee suggested that these dependencies are not causal but constitutive or due to part-whole relations. While there might be some part-whole relations involved in the model (e.g., in the way in which different processes contribute to the overall rate of change of a variable), the dependencies we are talking about here, e.g., the dependence of the rate of change of *M* on the concentration *M* (equation 1a) are not of this kind. This dependence is due to an enzyme-directed biochemical reaction, which is clearly a causal processes.

manipulated independently of all the other variables and equation as the standard causal theory requires it (see above; lack of independent manipulability and modularity). Rather, in these causal processes, concentrations and their rates of change are so intimately intertwined and integrated (cf. Mitchell 2009) that it is not possible to disentangle causal difference-making and dynamical information.

Why can the differential equations in the Goldbeter model not be replaced by something more akin to the causal modeler's structural equations, e.g., difference equations with a discrete time variable? As I have argued, it seems the same difficulty would arise: as soon as the concentration variables and the time intervals are fixed, the rates of change are determined and therefore no longer independent.[6] Furthermore, replacing the differential equations by standard structural equations would be like trying to do Newtonian mechanics without using calculus; what would be the point?

A possible response by the causal modeler might be to deny that the differential equations are even contenders for representing causal dependencies. Differential equations contain functions of time and their derivatives and need to be *integrated* in order to predict or explain physical events. Surely, when we want to discuss the causal content of models such as Goldbeter's we have to consider suitably *integrated* forms of equations.

The problem with this reply is that systems of differential equations such as Goldbeter's or HH can only be integrated numerically. The solutions of these equations that are available, showing the concentrations of various molecular species, have been obtained with the help of computer simulations. In these solutions, whatever causal

[6] This was pointed out to me by an anonymous referee.

difference-making information was represented in the differential equations (if any) is irretrievably lost. However, these simulations provide a different kind of information: They show under which parameter values certain kinds of behavior are stable. In the case of the Goldbeter model, the behavior that is of particular importance, for obvious reasons, is stable oscillatory behavior. It can be represented by a limit cycle in a plane defined by mRNA and total PER protein concentrations. The limit cycle gives the initial conditions for M and P_t (= total PER protein concentration) that generate a stable oscillation, which functions as the basic *Zeitgeber* for Drosophila's biological clock. I would not refer to this kind of information as causal difference-making information but as *stability information*.

Perhaps it could be argued that the integrated model provides some kind of causal difference-making information as well. In his original 1995 paper, Goldbeter showed that the rate of PER protein degradation has a strong effect on the period of the oscillations. The more rapidly the protein is degraded in the cell, the longer the period of the oscillations become. The reason is intuitively clear: The more rapidly the protein disappears, the longer it takes for protein synthesis to rise the concentration above the threshold where the repression of transcription by nuclear PER protein significantly slows down gene expression such that the concentration of PER starts to drop after a period of increase. However, as intuitively obvious as this may be, the exact effect of the rate of decay on the period of the oscillations can only be predicted by such a dynamical model, which, as I have shown, contains causal information that is highly *integrated* with temporal, dynamical information and thus not representable by standard causal models, because the independent manipulability and modularity requirements are not satisfied.

In the following section, I will consider some results from the causal modeling literature as to how systems of differential equations can be represented by structural causal models. As I will show, these results, while it is highly illuminating for the problem at hand, actually support my thesis about the limitations of causal modeling.

4. Differential Equations and Causal Structural Models

Attempts to describe at least the equilibrium states of systems of differential equations with structural causal models can be found in the causation literature, for example, Mooij, Janzing and Schölkopf (2013); henceforth abbreviated as "MJS".[7] MJS treat systems of ordinary first-order differential equations such as they feature in many scientific models, e.g., the Lotka-Volterra model of predator-prey dynamics or the coupled harmonic oscillator in mechanics. The systems described by such equations may be considered to contain causal cycles. For instance, in a predator-prey system the density of predators affects the density of prey, which causally feeds back to the predator density. This is the kind of causal cycle that we also find in our biological clock case examined in the previous section. Even though causal graphs (DAGs) are typically acyclic, this is not a constraint that would somehow be necessitated by the formalism. I have already mentioned possible approaches to modeling causal cycles in Section. MJS take a somewhat different approach: They show that the equilibrium solutions of systems of coupled differential equations that describe systems with some causal feed-back correspond to a structural causal model.

[7] I wish to thank an anonymous referee for calling this work to my attention.

It is not possible here to reproduce the full treatment given by MJS. Basically, they consider dynamical systems represented by systems of coupled differential equations of the following form:

$$\dot{X}_i(t) = f_i\left(X_{pa_{\mathcal{D}}(i)}\right), X_i(0) = (\mathbf{X}_0)_i \;\; \forall i \in \mathcal{I}$$

where the indices $pa_{\mathcal{D}}(i)$ range over the set of parents of the variable X_i, each f_i is a smooth function of X, and each $(X_0)_i$ is an initial condition. Then, they provide an account of what it means to *intervene* on such a system, as intervention is part of the standard semantics of causal models. In a nutshell, an idealized intervention can be described as:

$$\dot{X}_i(t) = \begin{cases} 0, & i \in I \\ f_i\left(X_{pa_{\mathcal{D}}(i)}\right), & i \in \mathcal{I} \setminus I \end{cases}$$

$$X_i(0) = \begin{cases} \xi_i, & i \in I \\ X_{0_i}, & i \in \mathcal{I} \setminus I \end{cases}$$

In such an intervention, some set of components I of the system are forced to take some target value, such that the first time derivative of the variable X_i takes the value zero (i.e., X remains constant), while the variable takes some fixed target value ξ_i.[8] Thus, whatever

[8] Note how this intervention must fix the values for both the variables and their rate of change at the same time (cf. Section 3.4). This is exactly how the structural causal model obliterates information that is explanatorily relevant.

mechanism previously determined the value of the X_i, the intervention exogenously breaks this mechanism and sets the variables to a fixed value. This corresponds to the well-known breaking of directed edges by intervention variables in ordinary causal graphs.

What is interesting to note in the present context is that, according to the definition of an idealized intervention given by MJS, such an intervention changes not just one but two equations. This shows, once again, that such a system of equations is not modular in the sense discussed in the previous section. (It might be modular in the sense that it doesn't change any further equations, though).

The idealized interventions obviously change the equilibrium states of the system. For example, a Lotka-Volterra system has a steady state in which the predator and prey populations show an undamped oscillation. If intervened upon in the manner shown above, such a system changes its equilibrium state. If, for example, the intervention sets the predator density in a Lotka-Volterra system to ξ_2, the system's new unique stable equilibrium state is $(X^{eq}_1, X^{eq}_2)=(0, \xi_2)$. In general, equilibrium states of systems of intervened differential equations can always be obtained by setting the rates of change of the variables to zero by an intervention. The resulting equilibrium is then described by some equilibrium equations.

Just as in ordinary causal graph representations, nodes and directed edges can be used to represent the outcome of possible interventions on the variables figuring in systems of differential equations. In the cases such as the ones considered here, there will be a set of equilibrium equations for each possible intervention of the kind introduced above. MJS show how such equilibrium equations can be derived in general, and that they form causal

structural models in accordance with the causal framework assumed. Thus, it seems that the causal graph framework with its standard interventionist semantics is able to deal with systems of differential equations. MJS suggest that this approach "sheds more light on the concept of causality as expressed within the framework of Structural Causal Models, especially for cyclic models."

I wish to draw a different conclusion from MJS's highly illuminating treatment. In my view, their approach to dynamical systems described by differential equations reveals precisely the limitations to the causal graph framework that I wish to expose in this paper. For it is clear that such an approach can only deal with stable equilibrium states of a system, i.e., such equilibrium states where there is no more change. This is a simple consequence from the kind of interventions introduced, where the first derivatives with respect to time of the variables considered are set to zero. Thus, the structural causal models represent static situations rather than dynamic processes. For some intents and purposes, this may be fine. But if it is accepted that the dynamical models examined here are representations of the causal properties of a system and that the rates by which variables change is such a property, this kind of causal property does not seem to be captured by ordinary causal structural models.

I wish to end this argument by disenabling a potential misinterpretation. My thesis of this paper should not be understood as a claim about causal *discovery*. None of the considerations presented here support the conclusion that a causal search procedure of the kind developed by Spirtes, Glymour and Scheines (2000) would be unable to identify all

the variables that values of which affect the behavior of the system.[9] I am only claiming that the entities referred to by these variables have causal properties – in particular the rate of change – that cannot be given a causal interpretation by using the standard formalisms.

6. Conclusions

My intention in this paper has not been to argue that there exist forms of explanations in biology that are not causal. There clearly is a sense in which DNA sequence recognition by proteins (Kaiser, this symposium) as well as the biological clock mechanisms discussed here are causal processes. What I as well as Kaiser (this symposium) want to show is that these biological explanations contain *causal* information that is not reducible to causal difference-making information of the kind that can be expressed in the formal causal models available today. Biological explanations often contain causal information that is inextricably intertwined with, first, *spatial* information and, second, *dynamical* information. The spatio-temporal aspects represented in these explanations are not such that they could simply be integrated with the causal difference-making information to give the full picture. At least in the case of the dynamical information contained in systems of differential equations, there appears to be a deep incompatibility between the axioms of causation and the dynamical model. Just as the circadian clock mechanism cannot be

[9] For an illuminating discussion of this important issue in the context of systems of differential equations see Dash (2005). It should be noted that, just like in the Mooij, Janzing and Schölkopf (2013), time derivatives of variables are never treated as independent causes.

understood by looking at the level of individual molecules, complex spatially organized cohesive interactions in DNA-protein complexes (see Kaiser, this symposium) cannot in practice be expressed by causal graphs in a way that brings out the explanatory power and utility of these models.

My conclusion with respect to dynamical mechanistic models differs thus somewhat from Bechtel's and Abrahamsen's illuminating analysis: What they call the sequential and dynamical mechanisms, respectively, represent not two models that complement each other. Rather, in my view they represent incompatible perspectives on the same phenomenon of the kind that scientific pluralists have postulated (Kellert, Longino and Waters 2006).

Thus, rather than just the need of supplementing causal graphs with spatio-temporal labels such as to fine-tune them, a close examination of biological explanations rather reveals some intrinsic limitations of a certain type of causal model. Perhaps a new theory of causation is needed in order to do (more) justice to such explanations, in biology as well as in other sciences that deal with complex dynamical processes.

References

Bechtel, William (2005), *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, William (2013), "From Molecules to Networks: Adoption of Systems Approaches in Circadian Rhythm Research", in Hanne Andersen, Dennis Dieks, Wenceslao J.

Gonzalez, Thomas Uebel and Gregory Wheeler (eds.), *New Challenges to Philosophy of Science*, Berlin: Springer.

Bechtel, William, and Adele Abrahamsen (2010), "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science", *Studies in History and Philosophy of Science Part A* 41:321-333.

Braillard, Pierre-Alain (2010), "Systems Biology and the Mechanistic Framework", *History and Philosophy of the Life Sciences* 32:43-62.

Casini, Lorenzo, Phyllis McKay Illary, Federica Russo, and Jon Williamson (2011), "Models for Prediction, Explanation and Control: Recursive Bayesian Networks", *Theoria. An International Journal for Theory, History and Foundations of Science* 70 (1):5-33.

Clarke, Brendan, Bert Leuridan and Jon Williamson (2014), "Modelling Mechanisms With Causal Cycles", *Synthese* 191 (8):1651-1681.

Craver, Carl (2007), *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Dupré, John (2013), "Mechanism and Causation in Biology. I—Living Causes", *Aristotelian Society Supplementary Volume* 87:19–37.

Gebharter, Alexander, and Marie I. Kaiser (2014), "Causal Graphs and Biological Mechanisms", in Marie I. Kaiser, O. Scholz, D. Plenge and A. Hüttemann (eds.), *Explanation in the Special Sciences. The Case of Biology and History*, Berlin: Springer.

Gebharter, Alexander (2014), "A Formal Framework for Representing Mechanisms?", *Philosophy of Science* 81 (1):138-153.

Goldbeter, Albert (1995), "A Model for Circadian Oscillations In the Drosophila Period Protein (PER)", *Proceedings of the Royal Society of London. B: Biological Sciences* 261 (1362):319-324.

Hardin, P. E., Hall, J. C., and M. Rosbash (1990), "Feedback of the *Drosophila Period* Gene Product On Circadian Cycling of Its Messenger RNA Levels", *Nature* 343 (6258): 536-540.

Hitchcock, Christopher, and James Woodward (2003), "Explanatory Generalizations, Part II: Plumbing Explanatory Depth", *Noûs* 37 (2):181-199.

Kellert, Stephen H., Helen E. Longino, and C. Kenneth Waters, eds. (2006), *Scientific Pluralism*, *Minnesota Studies in Philosophy of Science, Vol. XIX*. Minneapolis: University of Minnesota Press.

Kistler, Max (2013), "The Interventionist Account of Causation and Non-Causal Association Laws", *Erkenntnis* 78:1-20.

Kuhlmann, Meinard (2011). Mechanisms in Dynamically Complex Systems. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences*. Oxford: Oxford University Press.

May, R. M. (1974), "Biological Populations with Nonoverlapping Generations: Stable Points, Stable Cycles and Chaos", *Science* 186:645-647.

McKay Illari, P., Russo, F., & Williamson, J. (Eds.) (2011). *Causality in the Sciences*. Oxford: Oxford University Press.

Mooij, Joris M., Dominik Janzing, and Bernhard Schölkopf (2013), "From Ordinary Differential Equations to Structural Causal Models: The Deterministic Case", in Ann Nicholson and Padhraic Smyth (eds.), *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, Corvallis: AUAI Press, 440-448.

Otsuka, J. (forthcoming), "Causal Foundations of Evolutionary Genetics", *The British Journal for the Philosophy of Science*

Pearl, Judea (2000), *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Russo, Federica (2012), "Public Health Policy, Evidence and Causation: Lessons From the Studies On Obesity", *Medicine, Health Care and Philosophy* 15 (2):141-151.

Spirtes, Peter, Clark Glymour and Richard Scheines (2000), *Causation, Prediction, and Search*. Cambridge, Mass.: MIT Press.

Waskan, Jonathan (2011), "Mechanistic Explanation at the Limit", *Synthese* 183 (3):389-408.

Waters, C. Kenneth (2007), "Causes That Make a Difference", *The Journal of Philosophy* CIV (11):551-579.

Weber, Marcel (2005), *Philosophy of Experimental Biology*. *Cambridge Studies in Biology and Philosophy*. Cambridge: Cambridge University Press.

Weber, Marcel (2008), "Causes without Mechanisms: Experimental Regularities, Physical Laws, and Neuroscientific Explanation", *Philosophy of Science* 75:995-1007.

Weber, Marcel (2012), "Experiment in Biology", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/biology-experiment/

Woodward, James (2003), *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2008). Cause and Explanation in Psychiatry: An Interventionist Perspective. In K. S. Kendler, & J. Parnas (Eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. Baltimore: Johns Hopkins University Press.

Woodward, J. (2010). Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation. *Biology and Philosophy, 25,* 287-318.

Woodward, James (2011), "Mechanisms Revisited", *Synthese* 183 (3):409-427.

Woodward, James (2013), "Mechanistic Explanation: Its Scope and Limits", *Proceedings of the Aristotelian Society Supplementary Volume* 87 (1):39-65.

Woodward, James, and Christopher Hitchcock (2003), "Explanatory Generalizations, Part I: A Counterfactual Account", *Noûs* 37 (1):1-24.

Rethinking Boltzmannian Equilibrium

Charlotte Werndl and Roman Frigg
Department of Philosophy, Logic and Scientific Method
London School of Economics
c.s.werndl@lse.ac.uk, r.p.frigg@lse.ac.uk
Forthcoming in: *Philosophy of Science*

Abstract

Boltzmannian statistical mechanics partitions the phase space into macroregions, and the largest of these is identified with equilibrium. What justifies this identification? Common answers focus on Boltzmann's combinatorial argument, the Maxwell-Boltzmann distribution, and maximum entropy considerations. We argue that they fail and present a new answer. We characterise equilibrium as the macrostate in which the system spends most of its time and prove a new theorem establishing that equilibrium thus defined corresponds to the largest macroregion. Our derivation is completely general in that it does *not* rely on assumptions about the system's dynamics or internal interactions.

1 Introduction

Boltzmannian statistical mechanics (BSM) partitions the phase space of a system into cells consisting of macroscopically indistinguishable microstates. These cells correspond to the macrostates, and the largest cell is singled out as the equilibrium macrostate. The connection is not conceptual: there is nothing in the *concept* of equilibrium tying equilibrium to the largest cell. So what justifies the association of equilibrium with the largest cell?

After introducing BSM (Section 2), we discuss three justificatory strategies based on Boltzmann's combinatorial argument, the Maxwell-Boltzmann distribution, and maximum entropy considerations, respectively. We argue all three fail because they either suffer from internal difficulties or are restricted to systems with negligible interparticle forces. This prompts the search for an alternative answer. In analogy with the standard thermodynamic definition of equilibrium, we characterise equilibrium as the macrostate in which the system spends most of its time.

We then present a new mathematical theorem proving that such an equilibrium macrostate indeed corresponds to the largest cell (Section 4). This result is completely general in that it does it is not based on any assumptions either about the system's dynamics or the nature of interactions within the system.

2 Boltzmannian Statistical Mechanics

Let us briefly introduce BSM.[1] Consider a system consisting of n particles which is isolated from the environment and in a bounded container. The system's state is specified by a point $x = (q, p)$ (the *microstate*) in its $6n$-dimensional phase space Γ. The system's dynamics is determined by its classical Hamiltonian $H(x)$. Energy is preserved and therefore the motion is confined to the $6n - 1$ dimensional energy hypersurface Γ_E defined by $H(x) = E$, where E is the energy value. The solutions of the equations of motion are given by the phase flow ϕ_t on Γ_E, where $\phi_t(x)$ is the state into which $x \in \Gamma_E$ evolves after t time steps. Σ_E is the Lebesgue-σ-algebra and, intuitively speaking, consists of all subsets of Γ_E of interest. Γ is endowed with the Lebesgue measure μ, which is preserved under ϕ_t. This measure can be restricted to a measure μ_E on Γ_E which is preserved as well and is normalised, i.e. $\mu_E(\Gamma_E) = 1$. $(\Gamma_E, \Sigma_E, \mu_E, \phi_t)$ is a *measure-preserving dynamical system*.

Assume that the system can be characterised by a set $\{v_1, ..., v_k\}$ of *macro-variables* ($k \in \mathbb{N}$). The v_i assume values in $\mathbb{V}_i$, and capital letters V_i denote the values of v_i. A particular set of values $\{V_1, ..., V_k\}$ defines a *macrostate* $M_{V_1,...,V_k}$. We only write 'M' rather than '$M_{V_1,...,V_k}$' if the specific V_i do not matter. A set of macrostates is complete iff (if and only if) it contains all states a system can be in.

A crucial posit of BSM is supervenience: a system's microstate uniquely determines its macrostate. Every macrostate M is associated with a *macroregion* Γ_M consisting of all $x \in \Gamma_E$ for which the system is in M. For a complete set of macrostates the Γ_M form a partition of Γ_E (they do not overlap and jointly cover Γ_E).[2]

The *Boltzmann entropy* of a *macrostate* M is $S_B(M) := k_B \log[\mu(\Gamma_M)]$ (k_B is the Boltzmann constant). The Boltzmann entropy of a *system* at time t, $S_B(t)$, is the entropy of the system's macrostate at t: $S_B(t) := S_B(M_{x(t)})$, where $x(t)$ is the system's microstate at t and $M_{x(t)}$ is the macrostate supervening on $x(t)$.

We denote the equilibrium macrostate by M_{eq} and its macroregion by $\Gamma_{M_{eq}}$. A crucial aspect of the standard presentation of BSM is that Γ_{eq} takes up most of Γ_{M_E}. To facilitate the discussion, we introduce the term of β-dominance: $\Gamma_{M_{eq}}$ is

[1] For details see Frigg (2008, 103–21).

[2] For lack of space our focus is on the most common case where macrostates are defined relative to Γ_E. Our arguments generalise to cases when the macrostates are defined relative to other subsets of Γ.

β-dominant iff $\mu(\Gamma_{M_{eq}}) \geq \beta$ for $\beta \in [1/2, 1]$. Often equilibrium is characterised as a state where β is close to one but nothing in what follows depends on a particular choice of β.

The characterisation of equilibrium as a β-dominant state goes back to Ehrenfest and Ehrenfest (1912, 30). While different versions of BSM explain the approach to equilibrium differently, β-dominance is a key factor in all of them. Those who favour an explanation based on ergodic theory have to assume that $\Gamma_{M_{eq}}$ takes up the majority of Γ_E because otherwise the system would not spend most of the time in $\Gamma_{M_{eq}}$ (e.g. Frigg and Werndl 2011, 2012). Those who see the approach to equilibrium as the result of some sort of probabilistic dynamics assume that $\Gamma_{M_{eq}}$ takes up most of Γ_E because they assign probabilities to macrostates that are proportional to $\mu(\Gamma_M)$ and equilibrium comes out of as the most likely state only if the equilibrium macroregion is β-dominant (e.g. Boltzmann 1877). Proponents of the typicality approach see dominance as the key ingredient in explaining the approach to equilibrium and sometimes even seem to argue that systems approach equilibrium *because* the equilibrium region takes up nearly all of phase space (e.g. Goldstein and Lebowitz 2004).

We do not aim to adjudicate between these different approaches. Our question is a more basic one: *Why is the equilibrium state β-dominant*?

3 Justificatory Strategies

A look at the literature reveals three justificatory strategies. In practice these are often pursued side-by-side and seen as providing mutual support to each other. We will assess each of them and argue that none of them is conclusive.

3.1 The Largest Number of Microstates and the Combinatorial Argument

The leading idea of the first justificatory strategy is that *equilibrium is the macrostate that is compatible with the largest number of microstates*. This strategy is exemplified by Boltzmann's (1877) combinatorial argument.[3] The state of one particle is determined by a point in its 6-dimensional state space Γ_μ, and the state of a system of n identical particles is determined by n points in this space. Since the system is confined to a finite container and has constant energy E, only a finite part of Γ_μ is accessible. Boltzmann partitions the accessible part of Γ_μ into cells of equal size $\delta\omega$ whose dividing lines run parallel to the position and momentum axes. The result is a finite partition $\Omega := \{\omega_1, ..., \omega_m\}$, $m \in \mathbb{N}$. The cell in which a particle's state lies

[3]For details see Frigg (2008) and Uffink (2007).

is its *coarse-grained microstate*. The coarse-grained microstate of the entire gas, called an *arrangement*, is given by a specification of the coarse-grained microstate of each of particle.

The system's macro-properties depend only on how many particles there are in each cell and not on which particles these are. A specification of the 'occupation number' of each cell is known as a *distribution* $D = (n_1, n_2, \ldots, n_m)$ where n_i is the number of particles whose state is in cell ω_i. Since m and N are finite, there are only finitely many distributions $D_1, \ldots, D_k$. Each distribution is compatible with several arrangements, and the number $G(D)$ of arrangements compatible with a given distribution $D = (n_1, n_2, \ldots, n_m)$ is

$$G(D) = \frac{N!}{n_1! n_2! \ldots, n_m!}. \tag{1}$$

Every microstate x of Γ_E is associated with exactly one distribution $D(x)$. One then defines the set Γ_D of all x that are associated with a distribution D:

$$\Gamma_D = \{x \in \Gamma_E \ : \ D(x) = D\}. \tag{2}$$

Since macro-properties are fixed by the distribution, distributions are associated with macrostates. So we ask: which of the distributions is the equilibrium distribution? Now Boltzmann's main idea enters the scene: equilibrium is the macrostate that is compatible with the largest number of microstates. To determine the equilibrium distribution, Boltzmann assumed that the energy e_i of particle i depends only on the cell in which it is located. Then the total energy is:

$$\sum_{i=1}^{m} n_i e_i = E. \tag{3}$$

He furthermore assumed that the number of cells in Ω is small compared to the number of particles (allowing him to use Stirling's formula). With the further trivial assumption that $\sum_{i=1}^{m} n_i = N$, Boltzmann shows that $\mu_E(\Gamma_D)$ is maximal when

$$n_i = \gamma e^{\lambda e_i}, \tag{4}$$

where γ and λ are parameters which depend on N and E. This is the *discrete version of the Maxwell-Boltzmann distribution*. Thus the equilibrium macrostate corresponds to the Maxwell-Boltzmann distribution.[4]

[4]What (4) gives us is the distribution with the largest number of microstates (for the Lebesgue measure) on the $6N$-dimensional shell-like domain Γ_{ES} specified by the condition (3). It does *not* give us the macroregion of maximal size (i.e., the distribution with the largest measure μ_E on the $6n-1$ dimensional Γ_E). As Ehrenfest and Ehrenfest (1959, 30) stress, the (not further justified) assumption is that the possible distributions and the proportion of the different distributions would not change if macrostates were instead defined on Γ_E.

Its ingenuity notwithstanding, the combinatorial argument faces a number of important problems. The first is that it only applies to systems of non-interacting particles (Uffink 2007, 976-7). It provides a reasonable approximation for systems with negligible interparticle forces, but any other system is beyond its scope. SM ought to be a general theory of matter and so this is a serious limitation.

The second problem is the lack of a conceptual connection between equilibrium in thermodynamics (TD) and the idea that the equilibrium macrostate is the one that is compatible with the largest number of microstates. In TD equilibrium is defined as the state to which isolated systems converge when left to themselves and which they never leave once they have reached it. This has very little, if anything, in common with the kind of considerations underlying the combinatorial argument. This is a problem for anyone who sees BSM as a reductionist enterprise. While the precise contours of the reduction of TD to SM remain controversial, we are not aware of any contributors who maintain radical anti-reductionism. Thus the disconnect between the two notions of equilibrium is a serious problem.

Two replies come to mind. The first points out that since $\Gamma_{M_{eq}}$ is the largest subset of Γ_E, systems approach equilibrium and spend most of their time in $\Gamma_{M_{eq}}$. This shows that the BSM definition of equilibrium is a good approximation to the TD definition. This is not true in general. Whether a system spends most of its time in the β-dominant $\Gamma_{M_{eq}}$ depends on the dynamics. If, for instance, the dynamics is the identity function, it is not true that a system out of equilibrium approaches equilibrium and spends most of its time there. The second reply points out that we know for independent reasons that the equilibrium distribution is the Maxwell-Boltzmann distribution. This argument will be discussed in the next subsection, and our conclusion will be guarded.

Finally, the combinatorial argument (even if successful) shows that the equilibrium macrostate is *larger than any other macrostate*. However, as Lavis (2005) points out, this does not imply that the equilibrium is β-dominant for values of β close to 1. There may be a large number of smaller macrostates who *jointly* take up a large part of Γ_E. So the combinatorial argument does in fact not show that equilibrium is β-dominant.

3.2 The Maxwell-Boltzmann Distribution

According to the next justificatory strategy, *a system is in equilibrium when its particles approximately satisfy the Maxwell-Boltzmann distribution* (e.g., Penrose 1989). The Maxwell-Boltzmann distribution $f(x_\mu)$ specifies the fraction of particles in the gas whose 6-dimensional position and momentum coordinates $x_\mu = (q_\mu, p_\mu)$ lie in the infinitesimal interval $(x_\mu,\, x_\mu + dx_\mu)$:

$$f(x_\mu) = \frac{\chi_V(q_\mu)\,(2\pi mkT)^{-3/2}}{\|V\|} \exp\left(- \frac{p_\mu^2}{2mk_BT}\right), \tag{5}$$

where T is the temperature, $\|V\|$ is the volume of the container, and $\chi_V(x)$ is 1 if q_μ is in the container and 0 otherwise.

This approach is misguided because the Maxwell-Boltzmann distribution is in fact the equilibrium distribution only for a limited class of systems, namely for systems consisting of particles with negligible interparticle forces. For particles with non-negligible interations different distributions correspond to equilibrium (Gupta 2002). Furthermore, for many simple models such as the Ising model (Baxter 1982) or the Kac-ring (Lavis 2008) the equilibrium macrostate also does not correspond to the Maxwell-Boltzmann distribution.

This is no surprise given that the two main derivations of the distribution in effect assume that particles are non-interacting. Boltzmann's (1877) derivation is based on equation (3), the assumption that the total energy is the sum of the energy of the individual particles. This is true only if the particles are non-interacting, i.e. for ideal gases. While many expect that the argument also goes through for dilute gases (where this assumption holds approximately), the argument fails for non-negligible interactions. In Maxwell's 1860 derivation (see Uffink 2007) the non-interaction assumption enters via the postulate that the probability distributions in different spatial directions can be factorised, which is true only if there is no interaction between particles.

For these reasons the Maxwell-Boltzmann distribution is the equilibrium distribution only for a limited class of systems and cannot be taken as a general definition of equilibrium.

3.3 Maximum Entropy

A third stragey *justifies dominance by maximum entropy considerations* along the following lines:[5] we know from TD that, if left to itself, a system approaches equilibrium, and equilibrium is a maximum entropy state. Hence the Boltzmann entropy of a macrostate S_B is maximal in equilibrium. Since S_B is a monotonic function, the macrostate with the largest Boltzmann entropy is also the largest macrostate, which is the desired conclusion.

There are serious problems with the understanding of TD in this argument as well as with its implicit reductive claims. First, that a system, when left to itself, reaches equilibrium where entropy is maximal is often taken to be a consequence of

[5]This strategy has been mentioned to us in conversation but it is hard to track down in print, at least in pure form. Albert's (2000) considerations concerning entropy seem to gesture in the direction of the third strategy. See Winsberg (2004) for a detailed discussion of Albert's approach.

the Second Law of TD, but it is not. As Brown and Uffink pointed out (2001), that systems tend to approach equilibrium has to be added as an independent postulate, which they call the 'Minus First Law'. But even if TD is amended with the Minus First Law, the conclusion does not follow. TD does not attribute an entropy to systems out of equilibrium. Thus, characterising the approach to equilibrium as a process of entropy increase *is meaningless from a TD point of view*!

Even if all these issues could be resolved, there still would be a question why the fact that the TD entropy reaches a maximum in equilibrium would imply that the same holds for the Boltzmann entropy. To justify this inference, one would have to assume that the TD entropy reduces to the Boltzmann entropy. But it is far from clear that this is so. A connection between the TD entropy and the Boltzmann entropy has been established only for ideal gases, where the Sackur-Tatrode formula can be derived from BSM which shows that both entropies have the same functional dependence on thermodynamic state variables. No such results are known for systems with iteractions. Furthermore, there are well-known differences between the TD and the Boltzmann entropy. Most importantly, the TD entropy is extensive while the Boltzmann entropy is not (Ainsworth 2012). But an extensive concept cannot reduce to a non-extensive concept (at least not without further qualifications).

For these reasons we conclude that maximum entropy considerations cannot be used to argue for the β-dominance of the equilibrium state.

4 Rethinking Equilibrium

The failure of standard justificatory strategies prompts the search for an alternative answer. In this section we propose an alternative definition of equilibrium and introduce a new mathematical theorem proving that the equilibrium state is β-dominant.

The above strategies run into difficulties because there is no clear connection between the TD definition of equilibrium and β-dominance. Our aim is to provide the missing connection by taking as a point of departure the standard TD definition of equilibrium and then exploiting supervenience to 'translate' this macro definition into micro language.

The following is a typical TD textbook definition of equilibrium: 'A thermodynamic system is in equilibrium when none of its thermodynamic properties are changing with time [...]' (Reiss 1996, 3) In more detail: equilibrium is the state to which an isolated system converges when left to its own and which it never leaves once it has been reached (Callender 2001; Uffink 2001). Equilibrium in TD is unique in the sense thatf the system always converges toward the same equilibrium state. This leads to the following definition (the qualification 'strict' will become

clear later):

> *Definition 1: Strict TD Equilibrium.* Consider an isolated system S whose macrostates are specified in terms of the macro-variables $\{v_1, ..., v_k\}$. Assume the S is in state $M_{V_1,...,V_k}$ at some initial time t_0. If there is a macrostate $M_{V_1^*,...,V_k^*}$ satisfying the following condition, then it is the equilibrium state of S: For *all* initial states $M_{V_1,...,V_k}$ there exists a time t^* such that $v_i(t) = V_i^*$ for all all $t \geq t^*$, $i = 1, ..., k$. We then write $M_{eq} := M_{V_1^*,...,V_k^*}$.[6]

Note that this definition incorporates the Minus First Law of TD.

Rephrasing Definition 1 in the framework of BSM leads to the following definition:

> *Definition 2: Strict BSM Equilibrium.* Consider the same system S as above, now described as the measure-preserving dynamical system $(\Gamma_E, \Sigma_E, \mu_E, \phi_t)$, and let $M(x)$ be the macrostate that supervenes on microstate $x \in \Gamma$. Let $\Gamma_{M_{V_1,...,V_k}} := \{x \in \Gamma_E : M(x) = M_{V_1,...,V_k}\}$ be the set of all microstates on which $M_{V_1,...,V_k}$ supervenes. If there is a macrostate $M_{V_1^*,...,V_k^*}$ satisfying the following condition, then it is the strict BSM equilibrium state of S: For *all* initial states $x \in \Gamma_E$ there exists a time t^* such that $M_{V_1,...,V_k}(\phi_t(x)) = M_{V_1^*,...,V_k^*}$ for all $t \geq t^*$, $i = 1, ..., k$. We then write $M_{eq} := M_{V_1^*,...,V_k^*}$.

Before reflecting on this definition, we want to add a brief comment about reductionism. Reductive eliminativists may feel that a definition of equilibrium in SM that is based on 'top down translation' of its namesake in TD undermines the prospect of a reduction of TD to SM. They would argue that equilibrium has to be defined in purely mechanical terms, and must then be shown to line up with the TD definition of equilibrium.[7]

This point of view is not the only game in town and reduction can be had even if equilibrium is defined 'top down' (as in the above definition). First, whether the above definition undercuts a reduction depends on one's concept of reduction. For someone with a broadly Nagelian perspective there is no problem: the above definition provides a bridge law, which allows the derivation of the requisite macro regularities from the laws of the micro theory. And a similar argument can be made in the framework of New Wave Reductionism. Second, equilibrium is a macro concept: when describing a system as being in equilibrium, we look at it in

[6] If one wants to avoid the t^*-dependence on the initial state, one can instead demand that there exists a time t^* such that $v_i(t) = V_i^*$ for all initial states $M_{V_1,...,V_k}$ and all $t \geq t^*$, $i = 1, ..., k$.

[7] For a discussion see Dizadji-Bahmani et al. 2010.

terms of macro properties. From a micro point of view there are only molecules bouncing around. They always bounce – there is no such thing as a relaxation of particle motion to an immutable state. Hence a definition of equilibrium in macro terms is no heresy.

Definition 2 is too rigid for two reasons. The first reason is Poincaré recurrence: as long as the 'M' in SM refers to a mechanical theory that conserves phase volume (and there is widespread consensus that this is the case),[8] any attempt to justify an approach to strict equilibrium in mechanical terms is doomed to failure. The system will at some point return arbitrarily close to its initial condition, violating strict equilibrium (Frigg 2008; Uffink 2007). The second reason is that such a justification is not only unattainable but also undesirable. Experimental results show that equilibrium is not the immutable state that classical TD presents us with because systems exhibit fluctuations away from equilibrium (Wang et al. 2002). Thus strict equilibrium is actually *unphysical*.

Consequently, strict definitions of equilibrium are undesirable both for theoretical and experimental reasons. So let us relax the condition that a system has to remain in equilibrium for all $t \geq t^*$ by the weaker condition that it has has to be in equilibrium most of the time:

> *Definition 3: BSM α-Equilibrium.* Consider the same system as in Definition 2. Let $f_{M,x}(t)$ be the fraction of time of the interval $[t_0, t_0+t]$ in which the system's state is in M when starting in initial state x at t_0, and let α be a real number in $[0.5, 1]$. If there is a macrostate $M_{V_1^*,\ldots,V_k^*}$ satisfying the following condition, then it is the α-equilibrium state of S: For *all* initial states $x \in \Gamma_E$, $f_{M_{V_1^*,\ldots,V_k^*},x}(t) \geq \alpha$ in the limit $t \to \infty$. We then write $M_{\alpha\text{-}eq} := M_{V_1^*,\ldots,V_k^*}$.

An obvious question concerns the value of α. Often the assumption seems to be that α is close to one. This is reasonable but not the only possible choice. But for our purposes nothing hangs on a particular choice of α and so we leave it open what the best choice would be.

One last step is needed to arrive at the definition of equilibrium suitable for BSM. It has been pointed out variously that in SM, unlike in TD, we should not expect *every* initial condition to approach equilibrium (see, for instance, Callender 2001). Indeed, it is reasonable to allow for a set of very small measure ε for which the system does not approach equilibrium:

> *Definition 4: BSM α-ε-Equilibrium.* Let S and $f_{M,x}(t)$ be as above. Let α be a real number in $[0.5, 1]$ and let $1 > \varepsilon \geq 0$ be a small real number, and let Y be a subset of Γ_E such that $\mu_E(Y) \geq 1 - \varepsilon$. If there is a

[8] Hamiltonian Mechanics falls within this class, but the class is much wider.

> macrostate $M_{V_1^*,...,V_k^*}$ satisfying the following condition, then it is the α-ε-equilibrium state of S: For *all initial states* $x \in Y$ $f_{M_{V_1^*,...,V_k^*},x}(t) \geq \alpha$ in the limit $t \to \infty$. We then write $M_{\alpha\text{-}\varepsilon\text{-}eq} := M_{V_1^*,...,V_k^*}$.

Let us introduce the characteristic function of Γ_M, $1_M(x)$: $1_M(x) = 1$ for $x \in \Gamma_M$ and 0 otherwise. Definition 4 implies that for all $x \in Y$:[9]

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(\phi_t(x))dt \geq \alpha. \tag{6}$$

An important assumption in this characterisation of equilibrium is that μ_E (and not some other measure) is the relevant measure. It is often argued that μ_E can be interpreted as a probability or typicality measure (Frigg and Hoefer 2013; Werndl 2013). The condition then says that the system's state spends more than a fraction α of its time in equilibrium with probability $1-\varepsilon$, or that typical initial conditions lie on trajectories which spend more than α of their time in equilibrium.

We contend that *the relevant notion of equilibrium in BSM is α-ε-equilibrium.* The central question then becomes: why is the α-ε-equilibrium state β-dominant? Definition 4 in no way prejudges this question: it says nothing about the size of $\Gamma_{M_{\alpha\text{-}eq}}$ nor does it in an obvious sense imply anything about it.

That $\Gamma_{M_{\alpha\text{-}eq}}$ is β-prevalent follows from the following theorem, which we prove in the Appendix:

> *Equilibrium Theorem*: If $\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}$ is an α-ε-equilibrium of system S, then $\mu(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}) \geq \alpha(1-\varepsilon)$.

We emphasise that the theorem is completely general in that no dynamical assumption is made (in particular it is not assumed that the system is ergodic). So the theorem also applies to strongly interacting systems such as solids and liquids.

The Equilibrium Theorem is the centre piece of our account. It shows in full generality that *if the system S has an α-ϵ-equilibrium, then the equilibrium state is β-dominant for $\beta \geq \alpha(1-\varepsilon)$.*[10] This provides the sought-after justification of the β-dominance of the equilibrium state.

The Equilibrium Theorem makes the conditional claim that *if* there is an α-ϵ-equilibrium, *then* $\mu(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}) \geq \alpha(1-\varepsilon)$. As with all conditionals, the crucial and often vexing question is whether, and under what conditions, the antecedent holds. Some systems do not have equilibria. For instance, if the dynamics is given by the identity function then no approach to equilibrium takes place, and the antecedent of the conditional is wrong. By contrast, epsilon-ergodicity allows

[9] This shows that Defintion 4 is closely related to Lavis' (2005, 255) characterisation of TD-likeness.

[10] It is assumed that ε is small enough so that $\alpha(1-\varepsilon) \geq 0.5$.

for an equilibrium state to exist (Frigg and Werndl 2011). This raises the question under which circumstances the antecedent is true, which is an important question for future research.

5 Conclusion

Boltzmannian statistical mechanics partitions the phase space of a system into cells of macroscopically indistinguishable microstates. These cells are associated with the system's macrostates, and the largest cell is identified with equilibrium. What justifies the association of equilibrium with the largest cell? We discussed three broad justificatory strategies that can be found in the literature: that equilibrium is the macrostate compatible with the largest number of microstates, that equilibrium corresponds to the Maxwell-Boltzmann distribution and that most states are characterised by that distribution, and that equilibrium is the maximum entropy state. We argued that none of them is successful. This prompted the search for an alternative answer. We characterised equilibrium as the macrostate in which the system spends most of its time and presented a new mathematical theorem proving that such an equilibrium state indeed corresponds to the largest cell. This result is completely general in that it is not based on any assumptions either about the system's dynamics or the nature of interactions within the system. It therefore provides the first fully general justification of the claim that the equilibrium state takes up most of the accessible part of the system's phase space.

Appendix: Proof of the Equilibrium Theorem

The proof appeals to the ergodic decomposition theorem (cf. Petersen 1983, 81), stating that for a dynamical system $(\Gamma_E, \Sigma_E, \mu_E, \phi_t)$ the set Γ_E is the disjoint union of sets X_ω, each equipped with a σ-algebra Σ_{X_ω} and a probability measure μ_ω, and ϕ_t acts ergodically on each $(X_\omega, \Sigma_{X_\omega}, \mu_\omega)$. The indexing set is also a probability space $(\Omega, \Sigma_\Omega, P)$, and for any square integrable function f it holds that:

$$\int_{\Gamma_E} f d\mu_E = \int_\Omega \int_{X_\omega} f d\mu_\omega dP. \tag{7}$$

Application of the ergodic decomposition theorem for $f = 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(x)$ yields:

$$\mu_E(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}) = \int_{\Gamma_E} 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(x) d\mu_E = \int_\Omega \int_{X_\omega} 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(x) d\mu_\omega dP. \tag{8}$$

For an ergodic system $(X_\omega, \Sigma_{X_\omega}, \mu_\omega, \phi_t)$ the long-run time average equals the phase average. Hence for almost all $x \in X_\omega$:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(\phi_t(x)) dt = \int_{X_\omega} 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(x) d\mu_\omega = \mu_\omega(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}} \cap X_\omega). \tag{9}$$

From requirement (6) and because ϕ_t acts ergodically on each $(X_\omega, \Sigma_{X_\omega}, \mu_\omega)$, for almost all $x \in X_\omega$, $X_\omega \subseteq Y$:

$$\alpha \leq \lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(\phi_t(x)) dt = \int_{X_\omega} 1_{M_{\alpha\text{-}\varepsilon\text{-}eq}}(x) d\mu_\omega. \tag{10}$$

Let us first consider the case $\varepsilon = 0$, i.e. $\mu_E(Y) = 1$. Here from equation (8):

$$\mu_E(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}) \geq \int_\Omega \alpha dP = \alpha. \tag{11}$$

Hence if $\epsilon \geq 0$, it follows from equation (8) that:

$$\mu_E(\Gamma_{M_{\alpha\text{-}\varepsilon\text{-}eq}}) \geq \alpha(1 - \varepsilon). \tag{12}$$

REFERENCES

Ainsworth, Peter Mark 2012. "Entropy in Statistical Mechanics." *Philosophy of Science* 79: 542-560.

Albert, David 2000. *Time and Chance*. Cambridge/MA and London: Harvard University Press.

Baxter, Rodney 1982. *Exactly Solved Models in Statistical Mechanics*. San Diego: Academic Press Limited.

Boltzmann, Ludwig 1877. "Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung resp. den Sätzen über das Wärmegleichgewicht." *Wiener Berichte* 76: 373-435.

Brown, Harvey and Jos Uffink 2001. "The Origins of Time-Asymmetry in Thermodynamics: The Minus First Law." *Studies in History and Philosophy of Modern Physics* 32: 525-538.

Callender, Craig 2001. "Taking Thermodynamics Too Seriously." *Studies in History and Philosophy of Modern Physics* 32: 539-553.

Dizadji-Bahmani, Foad, Frigg, Roman and Stephan Hartmann 2010. "Who Is Afraid of Nagelian Reduction?" *Erkenntnis* 73: 393-412.

Ehrenfest, Paul and Tatiana Ehrenfest 1959. *The Conceptual Foundations of the Statistical Approach in Mechanics*. Ithaca, New York: Cornell University Press.

Frigg, Roman 2008. A Field Guide to Recent Work on the Foundations of Statistical Mechanics. In *The Ashgate Companion to Contemporary Philosophy of Physics*, ed. Dean Rickles 99-196. London: Ashgate.

Frigg, Roman and Carl Hoefer 2010. Determinism and Chance from a Humean Perspective. In *The Present Situation in the Philosophy of Science*, ed. Dieks, Dennis, Wencelao Gonzales, Stephan Hartmann, Marcel Weber, Friedrich Stadler und Thomas Übel, 351-372. Berlin and New York: Springer.

Frigg, Roman and Charlotte Werndl 2012. "Demystifying Typicality." *Philosophy of Science* 79:917-929.

Frigg, Roman and Charlotte Werndl 2011. "Explaining Thermodynamic-Like Behaviour in Terms of Epsilon-Ergodicity." *Philosophy of Science* 78: 628-652.

Goldstein, Sheldon and Joel L. Lebowtiz 2004. "On the (Boltzmann) Entropy of Nonequilibrium Systems." *Physica D* 193:53-66.

Gupta, Mool C. 2003. *Statistical Thermodynamics*. New Delhi: New Age International Publishing.

Lavis, David 2005. "Boltzmann and Gibbs: An Attempted Reconciliation." *Studies in History and Philosophy of Modern Physics* 36: 245-273.

Lavis, David 2008. "Boltzmann, Gibbs and the Concept of Equilibrium." *Philosophy of Science* 75:682-696.

Penrose, Roger 1989. *The Emperor's New Mind*. Oxford: Oxford University Press.

Petersen, Karl 1983. *Ergodic theory*. Cambridge: Cambridge University Press.

Reiss, Howard 1996. *Methods of Thermodynamics*. Mineaola/NY: Dover.

Uffink, Jos 2001. "Bluff Your Way in the Second Law of Thermodynamics." *Studies in History and Philosophy of Modern Physics* 32: 305-394.

Uffink, Jos 2007. Compendium of the Foundations of Classical Statistical Physics. In *Philosophy of Physics*, ed. Jeremy Buttereld and John Earman, 923-1047. Amsterdam: North Holland.

Wang, Genmiao, Sevinck, Edith. M., Mittag, Emil, Searles, Debra J. and Denis J. Evans 2002. "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales." *Physical Review Letters* 89:050601.

Werndl, Charlotte 2013. "Justifying Typicality Measures in Boltzmannian Statistical Mechanics." *Studies in History and Philosophy of Modern Physics* 44: 470-479.

Winsberg, Eric 2004. "Can Conditioning on the Past Hypothesis Militate Against the Reversibility Objections?" *Philosophy of Science* 71: 489-504.

Evidential Critieria of Homology for Comparative Psychology

Abstract: While the homology concept has taken on importance in thinking about the nature of psychological kinds (e.g. Griffiths 1997), no one has shown how comparative psychological and behavioral evidence can distinguish between competing homology claims. I adapt the operational criteria of homology to accomplish this. I consider two competing homology claims that compare human anger with putative aggression systems of nonhuman animals, and demonstrate the effectiveness of these criteria in adjudicating between these claims.

Word count: 4826

1. Introduction

Many emotion researchers and theorists have suggested that anger is an innate adaptation that may be shared with nonhuman animals (e.g. Ekman 1999; Sell, Tooby, and Cosmides 2009). This raises the question of which behaviors might be manifestations of anger in non-human animals. Given the tight link between anger and aggression in humans, some aggression researchers propose that innate patterns of aggression in nonhuman animals are manifestations of anger. In other words, they propose that the system responsible for these phenomena is homologous with human anger, meaning that these complex traits are derived from a common ancestral trait.

As plausible as this may sound, there have been two incommensurate proposals along these lines, and there has been little progress in adjudicating between them. According to the *ethological hypothesis*, a repertoire of confrontational behaviors observed in "resident", territory-holding, rats reflects "an underlying emotional state" that is a primitive version of anger (Blanchard and Blanchard 1984, 17 see also; Blanchard and Blanchard 1988; Blanchard and Blanchard 2003). This behavioral repertoire is set in opposition to avoidance behaviors observed in intruder rats, which reflect fear. Moreover, the hypothesis holds that these two distinct emotional systems provide the best way of understanding angry aggression and fearful aggression in humans. Another proposal, the *neurophysiological hypothesis* is that human experiences of anger "emerge" from a pan-mammalian brain system that produces defensive behaviors that are elicited when areas within the ventral hypothalamus (among other areas) are electrically stimulated (Panksepp and Biven 2012; Panksepp 1998; Panksepp and Zellner 2004). These behaviors are set in opposition to predatory behaviors,

which are neurally dissociable from the defensive behaviors. In other words, this hypothesis holds that there are two neural systems for aggression, and that one of them, the defensive aggression system, provides the primary neural substrate for human anger and is the proximate cause of "the feeling states and behavioral acts" (Panksepp, 1998, p. 14) distinctive of human anger. Moreover, the proponents of this hypothesis claim that we can best understand certain types of human aggression, impulsive and instrumental forms of aggression, in terms of the neural systems for defense and predation, respectively.

Importantly, these hypotheses are incompatible. Within the neurophysiological tradition, the neural dissociation between predatory and defensive aggression is the main reason to consider them fundamental, distinct categories of aggression. However, confrontation and avoidance behaviors do not exhibit this kind of clean neural dissociation (Siegel 2004, chap. 1). Moreover, the kinds of defensive aggression in rats produced by electrical brain stimulation is distinct from the aggression observed in ethological research in the sense that it lacks features that are diagnostic of these forms of aggression (e.g. Kruk 1991). In other words, the aggression phenomena identified by these different research programs are behaviorally distinct and distinct neural mechanisms are responsible for them. As a result, they make incompatible inferences about what anger is and, more specifically, about which aggression phenomena are its manifestations. The bimodal classification schemes for aggression (defensive versus predatory and confrontational versus avoidant) that distinguish these respective phenomena are incommensurate.

While proponents of these hypotheses aim to identify homologies, there has been little progress in adjudicating between them. There are two reasons for this. One is the *target*

problem: they have not carefully identified the human psychological trait that is the target of comparison. Another is the *evidence* problem: it is unclear how cross-species comparisons support homology claims. More specifically, it remains obscure how comparative evidence can play a role in adjudicating competing homology claims. While the issues pertaining to the target problem have received a good deal of attention in philosophy of biology, the evidence problem has been neither raised nor resolved. In this paper, I show a way forward by developing evidential criteria of homology and an evidential constraint on homology claims. I then apply these criteria to the case of human anger and animal aggression to make it clear how hypothesis testing can proceed.

In the following section, I say more about homology thinking. Homology thinking is a historical mode of thinking that explains similarities by appealing to common descent. To understand what kind of evidence supports homology, I point out a range of hypotheses with which it competes and set out the kind of evidence that favors homology over and above them. The operational criteria of homology (Remane 1971) can be understood as identifying similarities that provide evidence for homology over and above these competing hypotheses. When the criteria are used in this way, I refer to them as the *evidential criteria of homology*. In section 2, I briefly address the target problem. Then I show how the evidential criteria of homology apply to the case of human anger and the aggression systems of nonhuman animals. A straightforward application of the criteria provides stronger support for the ethological hypothesis. Basic human anger has several similarities with the confrontational behaviors of resident rats, which provide some evidence that these traits are a product of common ancestry. On the other hand, there is currently no evidence that the defensive

aggression system uniquely corresponds with human anger. The similarities identified by the neurophysiological hypothesis hold not only with anger but also with other human emotions, such as fear. I conclude by highlighting the value of cross-species comparison for specifying psychological kinds.

2. Homology and its Competitors

Though the concept of homology is crucial to evolutionary thinking, it was conceived in the service of biological taxonomy prior to Darwin's time. Owen (1846) thought of homology as the sameness of an organ or structure in different organisms under every *form* and *function*. A common example of homology is the skeletal anatomy of the vertebrate forelimbs. The radius and ulna are bone structures that are common to bats, chimps, giraffes and manatees even though their forms and functions are dramatically different among these animals (see Figure 1). They can be more or less dense, thicker or thinner, longer or shorter, (though their spatial relationship to other bones of the forearms are preserved) and they can contribute to the different functions of swimming, flying, running and grasping in different organisms. So the radius and ulna are the same traits that occur in different animals, even though they have widely varying forms and functions within these various animals.

Now that evolutionary thinking has been integrated into biological systematics, one prominent idea about homology is that homology is a causal-historical concept (see Ereshefsky 2012 for a clarification and defense of this claim). Specifically, a homology refers to traits of various animals that derive from a trait of a common ancestor. In this way, shared ancestry is the common cause of each homologue, and this common cause explains

similarities between the homologous traits. In the words of one biologist (with some help from Darwin), homology is "…grounded in 'descent, with modification,' a process that belongs to the past." (Rieppel 2005, 24)

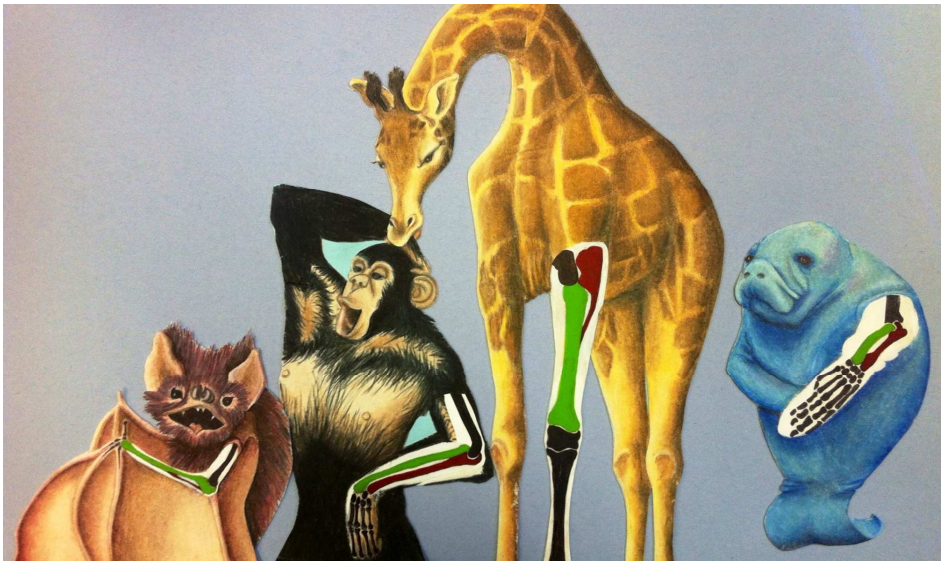


Figure 1. The bones of some mammalian forelimbs. The radius (green) and ulna (red) are the same kind of bone, which takes on different forms and functions in different animals.

As a causal-historical concept, we can identify and refer to a homology without having or requiring detailed knowledge of the developmental and hereditary mechanisms that give rise to it, just as we can refer to a disease entity, such as measles or chicken pox, without knowing about its underlying causes (Putnam 1969). Nonetheless, we learn more about each homology as we learn more about its underlying causes, just as we learn more about chicken pox as we learn more about the virus that causes it.

Given the causal-historical nature of homology, there is a vast range of evidence that could bear on whether or not one trait is homologous to another. Some of the best evidence pertaining to homology comes from cladistics. If one has an independently established phylogenetic tree, one can look at the distribution of a candidate homology, or character, on that tree. If, for instance, the existence of a homology is more parsimonious than convergent evolution on one or more occasion, then there is a strong reason to think that a trait is homologous.

Nevertheless, before we can even look at the distribution of a character on a phylogenetic tree, we need to know how to identify the character in each taxon, which becomes a tricky matter when dealing with behavioral and psychological characters. For instance, knowing that humans have anger, that rats have a confrontation system, and that cats have a defensive aggression system does not determine which of these capacities are the *same* trait or character.

One way of addressing this problem is to use the operational criteria of homology. These criteria need not function as a definition of homology but instead we can use them to establish a consistent set of methods for ascertaining homologies and by extension, identical traits. The criteria of homology attempt to identify particular kinds of similarity, the kinds that are best explained by common history over and above a range of competing hypotheses. For any given similarity across clades, there are several hypotheses in competition with homology. One is that the similarity is only by chance. Another more probable possibility is that convergent evolution explains the correspondence. When a similarity is explained purely by convergent evolution, we have a clear case of analogy. Still another possibility in the

behavioral domain is that similarity is explained by plastic developmental processes, particularly learning. In the clearest cases of plasticity, similarity can be explained entirely by convergent learning or development, perhaps shaped largely by task demands or shared developmental mechanisms.[1] The main competition is thus between hypotheses of homology, analogy, and developmental plasticity. Insofar as they function as evidence, the criteria of homology should help pick out similarites between traits that are explained by common ancestry and not convergent evolution or plastic developmental processes.

The most prominent criteria for homology were developed by Adolf Remane (1971) and can be deployed for this purpose. Consider first the criterion of *position*. The criterion applies to the radius and ulna because even with different forms and functions across different organisms, they retain their relative position to other bones of vertebrate forelimbs (humerus and the bones of the wrist). It would be highly unlikely for these characters to have evolved de novo in each of the different animals that possess it and yet to have the same relative position to other structures. Moreover, there is no shared function across the different animals which possess this character that would explain the correspondence. While corresponding position sounds like a spatial property, it is actually topological, and can include corresponding positions in temporal sequence or corresponding positions across cognitive architectures (e.g. "boxologies").

The criterion of *special quality* concerns "…shared features [that] cannot be explained by the role of a part in the life of the organism. The fact that in the vertebrate eye the blood supply to the retina lies between the retina and the source of light is a famous

[1] See Brown (2013) for a detailed discussion of the difficulties (e.g. due to the plasticity and transformability of behavior) in applying the criteria of homology to behavior.

example of a 'special quality'." (Griffiths 2007, 648) The more complex a shared quality is, the less likely that they would have evolved independently. The location of blood supply to the vertebrate retina is both complex and non-essential (and even slightly counterproductive) given the functional role of the retina (what it is used for in the organism), so it identifies a correspondence that provides strong evidence that the various instances of this character derive from common descent.

Finally, the criterion of *intermediate forms* allows identification of homologous forms, A and C, because of the existence of one or more transitional states, $B_1...B_n$, between the two forms. In many cases, the homology between transitional forms, say between A and B_1 or between B_1 and B_2, is determined by applying the other two criteria. For instance, there are transition states between the bones of the mammalian inner ear and the bones of the reptilian jaw. We know this because the bones of the reptilian jaw share the same *position* (relative to other bones of the jaw) as the bones of several intermediate forms, some of which share the same position as the bones of the mammalian inner ear.

For my purposes, an important constraint on homology claims derives from the fact that some homologies are nested within other homologies. For instance, the class of tetrapod forelimbs is nested within the class of paired appendages. Thus, the forelimbs of reptiles, amphibians, mammals and avians are members of the homology class of tetrapod forelimbs, but they are also members of the more inclusive homology class of paired appendages, which also includes the pectoral fins of sharks and teleosts. While pectoral fins are homologous with instances of tetrapod forelimbs *as paired appendages*, the similarities between pectoral fins and tetrapod forelimbs do not provide evidence for homology at the less inclusive level

of tetrapod forelimbs. Inclusion in this more specific class is indicated by bone structures that are absent in pectoral fins. These structures are due to modifications that occured subsequent to the divergence of tetrapods from teleosts, and that is why teleost pectoral fins are not included in this homology class.

As a result, some similarities only indicate inclusion in a broader homology class (e.g. paired appendages), whereas other similarities indicate inclusion in narrower homology classes (e.g. tetrapod forelimbs). In other words, some similarities (e.g. those between pectoral fins and forelimbs) only provide *evidence* for inclusion in broader homology classes (e.g. paired appendages rather than tetrapod forelimbs). It follows that, when evaluating similarities between traits, it is sometimes necessary to consider which homology class a similarity indicates.

From these considerations, we can derive an evidential constraint on homology claims. To see this, consider the correspondences between a human forelimb and a feline *hind limb*. The criterion of position is satisfied, because there are similarities between the parts (e.g. between humerus and femur). There *are* relations of homology between these traits. They are homologous as mammalian extremities and tetrapod extremities. Nevertheless, if we were to specify the homology class as one that includes human forelimbs but excludes human hind limbs, the similarity in question does not provide evidence for homology at this level.[2] This is because there are no similarities between the human forelimb and cat hind limb that *are not also shared between the human forelimb and hind limb*. Thus, to provide evidence for relations of homology at the level of some homology class G (in this

[2] For similar reasons, we also do not have evidence here for a relation of homology that would include feline hindlimbs but not feline forelimbs.

case, the homology class that includes forelimbs but excludes hindlimbs) *as opposed to the more inclusive class,* H (in this case, homology classes that includes forelimbs and hindlimbs), requires that some similarities between relata are not shared by traits in the more inclusive class, H. I call this an "evidential constraint" on homology claims.

While the examples so far deal straightforwardly with morphology or body structure, all three of Remane's criteria have also been applied to behavioral and psychological traits by ethologists (for overviews, see Ereshefsky 2007; Wenzel 1992). I suspect that what seems obvious concerning morphology might be easily confused concerning behavior or psychology. As a result, one could find evidence that psychological traits are homologous, but misidentify the homology class that this evidence supports. One way of doing so is to violate the evidential constraint above. I will argue that the neurophysiological hypothesis is an instance of this mistake. As yet, there is no evidence that the defensive aggression system identified by neurophysiological research is a member of the homology class that includes anger but excludes other human emotions. This is because the hypothesis does not identify any similarities that are not shared with other human emotions. I spell out the details of this argument in the following section.

In summary, homology is a causal-historical concept, and homology thinking is a way of providing historical explanations for observed similarities between biological traits (or characters). The evidential criteria for homology can isolate evidence pertaining to this kind of historical explanation. In the following section, I show how the evidential criteria can discriminate between the two hypotheses laid out in section 2.

3. Which Kinds of Aggression are Manifestations of Anger?

Before I apply the criteria to the two comparative hypotheses, I will first say something about the target problem, the problem of specifying the psychological trait that is the target of comparison. For the sake of space, I assume that the appropriate target of the ethological and neurophysiological hypotheses is *basic human anger*, the cluster of properties associated with involuntary facial expressions of human anger (Ekman 1999; Griffiths 1997). To briefly defend this choice, this is the most closely studied set of "anger" phenomena the structure of which is likely explained by inheritance, therefore it is the most plausible target for homology claims. This is because homology claims identify traits across taxa that are inherited from a common ancestor. One might say that inheritance is one of the causal homeostatic mechanisms that preserve the structure of homologous traits across lineages (cf. Assis and Brigandt 2009; Brigandt 2009). Thus, if there is something like anger in non-human animals, then it is most likely to correspond with phenomena in humans that are explainable by inheritance, namely basic human anger.

Now we are in a position to evaluate the two hypotheses. Recall that the two hypotheses focus on different sets of phenomena. The ethological hypothesis focuses on patterns of confrontational behavior of territory-holding, "resident" rats, whereas the neurophysiological hypothesis focuses on patterns of defensive behavior elicited by electrical brain stimulation. The ethological hypothesis lumps its phenomena together according to contrasting motives of behavior (confrontation versus defense), whereas the neurophysiological hypothesis lumps its phenomena together according to dissociable neural

substrates of behavior (regions of the hypothalamus that elicit defense behavior versus distinct regions that elicit predation behavior).

First, consider the ethological hypothesis. The strongest pieces of evidence for homology is a special quality that is shared by rats and stumptail macaques. Adams and Shoel (1981) note that dominant macaques and resident rats both implement strategies aimed at accessing the back and biting it. In macaques, this behavior seems arbitrary with respect to the (probable) function of inflicting non-lethal damage on the subordinate. Macaques have a much larger repertoire of bodily movements than rats, many of which could serve the function of inflicting non-lethal harm (pushing, kicking, scratching, slapping, holding etc.). Thus, back-biting is a *special quality*, and the best explanation of this behavior may appeal to products of common ancestry. In other words, the reason that the attacks of both rats and macaques are aimed at biting the neck and back may be that they share a common ancestor with a corresponding aggressive strategy and perhaps similar motivational mechanisms for negotiating intraspecific conflict.[3] There is some evidence that human anger includes an impulse to approach and attack, but no one has demonstrated that the impulse is pan-cultural or species-typical.[4]

While Adams and Schoel did observe several facial expressions of subordinate macaques, they did not note any facial expressions that uniquely accompanied the attacks of a dominant macaque. However, in more ecologically valid studies of macaque behavior, macaques with higher dominance status do display facial expressions toward lower ranking

[3] Adams and Schoel argue for homology by considering similarity in the dynamic of attack and submission across both species.
[4] See e.g. Carver and Harmon-Jones (2009); Baron (1971); Berkowitz et al (1981); and Pedersen et al (2011).

macaques in aggressive encounters, expressions that resemble anger expressions in humans (Chevalier-Skolnikoff 1974).[5] Chevalier-Skolnikoff (1973) argues that two of these expressions are similar (utilizing homologous action units) across macaques, chimps, and humans. Some confirmation of these comparisons has been attained by comparison using a facial action coding system to quantify chimpanzee facial expressions (Parr et al. 2007). Thus, there is *continuity across the intermediates* for some components of putative aggression systems across the common ancestors of these species.

Now consider the neurophysiological hypothesis. The problem is that the case for homology is incomplete. First, there is some evidence for correspondence that has *continuity across intermediates*: stimulation of the hypothalamus of cats, possums, rats and marmoset monkeys leads to similar forms of attack (Roberts, Steinberg, and Means 1967; Bergquist 1970; Panksepp 1971; Woodworth 1971; cited in Lipp and Hunsperger 1978).[6] However, ethical and practical considerations make it nearly impossible to obtain evidence concerning the effects of hypothalamus stimulation in humans. It remains uncertain whether it would leads to attack or to any of the other concomitants of human anger (e.g. experiences of anger, facial expressions of anger, or physiological changes associated with anger, as distinct from fear). Nor have any of these studies observed distinctive facial expressions that indicate continuity with human anger.[7]

---

[5] Chevalier-Skolnikoff calls these expressions "stare", "round-mouthed stare" and "open-mouthed stare".

[6] Delgado (1968) produced aggressive behaviors with electrical stimulation of the thalamus and cerebellum of chimpanzees and macaques. However, these brain structures are notably absent from the neurophysiological hypothesis and its descriptions of brain structures involved in aggression. Moreover, Delgado and colleagues did evaluate facial expressions. However, these facial expressions were not analyzed.

[7] It is compelling that in macaques, stimulation only results in attack under certain conditions (Alexander and Perachio 1973), some of which depend on whether the electrical stimulation occurs in the presence of a higher or lower ranking conspecific (attack being more likely in the latter case). Neverthless, one cannot conclude from

There is some evidence that amygdala stimulation can produce feelings of anger (e.g. Hitchcock and Cairns 1973). This evidence is even bolstered by the fact that stimulation of the medial amygdala in cats can potentiate defensive behaviors elicited by electrical stimulation of the hypothalamus (e.g. Shaikh, Steinberg, and Siegel 1993). However, several other emotional experiences beside anger have also been reported as a result of amygdala stimulation in humans, including anxiety, guilt, embarrassment, jealousy, and a "desire for flight or escape" (which is more strongly associated with human fear, see Frijda, Kuipers, and ter Schure 1989). It seems that current evidence does not support a distinct localization of anger-like and fear-like feelings or behaviors within the HAA or in the other brain structures that make up the defensive aggression system (in cats or otherwise). Thus, the evidence from brain stimulation does not reveal a unique correspondence with human anger; one that is not also shared with other human emotions.

Second, consider the criterion of position. As with the offensive attack observed in ethological work, physiological arousal and threat signals do occur prior to defensive attacks elicited by electrical brain stimulation. However, no evidence has been presented that either the signals or physiological arousal involved in these attacks are homologous with these

---

this that this form of aggression is of a piece with the aggressive syndrome which includes angry facial expression. It is quite possible that there are several forms of impulsive aggression that an animal might inflict only upon lower ranking conspecifics, including pain induced aggression, fear induced aggression or perhaps even disgust induced aggression. Neither is it obvious that any of these forms of aggression are of the same kind as angry aggression. By contrast, the work of Adams and Schoel (1981), and Chevalier-Skolnikoff (1973) describes a certain kind of offensive or dominance-related aggression *with which angry facial expressions are associated*. The same is not true of aggression elicited by electrical brain stimulation. The connection with angry facial expressions has not been made, nor has the behavioral syndrome been carefully circumscribed in ecologically valid conditions in most of the organisms in which it has been observed. Leyhausen (1979) has done this work concerning defensive aggression in cats, but he distinguishes this form of aggression from an offensive form of aggression that includes a back-biting attack. I suspect that this latter form of aggression is more comparable to the confrontation system in rats (cf. Blanchard and Blanchard 1984).

components of human anger as opposed to human fear. Moreover, it seems unlikely that any such evidence will materialize.

This becomes apparent when we look closely at the work of Siegel and others on the HAA, which is cited as support for the neurophysiological hypothesis (Panksepp 1998, 2012). In fact, Siegel does not advocate the neurophysiological hypothesis, and in many cases makes claims that constitute evidence against it. In several places (including Siegel 2004) Siegel compares defensive behaviors with a disorder known as Episodic Discontrol, which is marked by "...decreased impulse control – a characteristic common to defensive behavior – and altered perceptual states following stimuli evoking *anger, fear or rage*." (Siegel and Victoroff 2009, 213 emphasis mine) Indeed, many of the similarities that are noted between defensive behaviors and these forms of human aggression are characteristics of affectively driven behavior in general. Impulsivity is a characteristic of many kinds of emotion expression (see e.g. Frijda 1986), including fear, anger, sadness, and joy. Thus, the position criterion is not satisfied in a way that provides evidence for a homology between the defensive aggression system and anger that is not also shared between human anger and human fear.

By contrast, manifestations of the confrontation and avoidance systems in rats can be distinguished by quantifiable differences in the facial expressions of residents and intruders (Defensor et al. 2012), just as manifestations of anger and fear in humans can be distinguished by their distinctive facial expressions (e.g. Ekman and Friesen 1971). Moreover, resident and intruder rats have distinct forms of attack with distinct target sites. Thus, it is possible to distinguish *within rats* at least two different patterns of impulsive

behaviors accompanied by distinct facial expressions. Moreover, some of the similarities between confrontation behaviors and angry behaviors in humans are not shared with fearful behaviors in humans or avoidance behaviors in rats. In other words, human anger and the confrontation system in rats do not violate the evidential constraint on homology claims (relativized to a homology class that only includes the emotion of anger) because they satisfy the evidential criteria of homology in ways that are not also satisfied by other emotions like fear. A related virtue of the ethological hypothesis is that it can distinguish angry aggression from the widely acknowledged category of *fear-induced* aggression (see esp. Moyer 1976). The same cannot be said for the neurophysiological hypothesis. I suspect that at least some of the phenomena identified by the neurophysiological hypothesis reflect behavioral outcomes of fear, rather than (or perhaps in addition to) anger.

In sum, the case for homology between the defensive aggression system and anger (with respect to a category that includes anger but not other human emotions) may be similar to the case for homology between the cat hind limb and the human forelimb (with respect to a category that includes human forelimbs but not human hind limbs). The similarities so far observed do not evince a homology relation that excludes other emotions (especially fear), whereas the case for homology between the offensive attack system and anger does evince such a relation.

4. Conclusion

I have argued that the available evidence supports a homology between human anger and the confrontational attack system and not between human anger and the defensive

aggression system. However, this case study has larger implications for the scientific study of psychological kinds. The lesson is this: homology thinking can provide independent criteria for evaluating substantive disagreements on – and for eliminating confusion about – the nature of psychological kinds. In absence of homology thinking, it is difficult to see how further knowledge about the defensive aggression system or the offensive attack system would serve to determine which aggression systems in non-human animals are most like human anger. Indeed, this is probably one of the reasons why there has been little productive discussion between the advocates of the two hypotheses. Homology thinking in this case provides a set of independent theoretical constraints for identifying corresponding traits across taxa. In the service of this demonstration, I further developed some of the methods of homology thinking (Ereshefsky, 2007, 2012) as it applies to psychological kinds. This account helps to specify what kind of evidence supports homology claims, namely, identification of *unique* correspondences *at the appropriate level* between traits; correspondences that provide evidence for common ancestry as opposed to common selective pressures (whether developmental or ancestral).

Though counterintuitive from some perspectives, the concept of homology helps to clarify what counts as evidence for claims of trait identity. Note that identical traits can have different states. For example, a human arm and whale fin are identical traits, because they are both instances of the tetrapod forelimb. Nevertheless, they are different states of that trait, because they represent different forms that this trait can take. Homology thinking allows the identification of traits that take shape in dramatically different states; it enables us to identify evolved characters that walk in the guise of dramatically different forms and functions.

Anger is one such character.

References

Adams, David B. 1981. "Motor Patterns and Motivational Systems of Social Behavior in Male Rats and Stumptail macaques—Are They Homologous." *Aggressive Behavior*.

Alexander, M., and A. A. Perachio. 1973. "The Influence of Target Sex and Dominance on Evoked Attack in Rhesus Monkeys." *American Journal of Physical Anthropology* 38: 543–548.

Assis, Leandro C. S., and Ingo Brigandt. 2009. "Homology: Homeostatic Property Cluster Kinds in Systematics and Evolution." *Evolutionary Biology* 36 (2) (March 19): 248–255. doi:10.1007/s11692-009-9054-y.

Baron, Robert A. 1971. "Magnitude of Victim's Pain Cues and Level of Prior Anger Arousal as Determinants of Adult Aggressive Behavior." *Journal of Personality and Social Psychology* 17 (3): 236–243. doi:10.1037/h0030595.

Bergquist, E. H. 1970. "Output Pathways of Hypothalamic Mechanisms for Sexual, Aggressive and Other Motivated Behaviors in Oppossum." *Journal of Comparative and Physiology and Psychology* 70: 389–398.

Berkowitz, Leonard, S T Cochran, and M C Embree. 1981. "Physical Pain and the Goal of Aversively Stimulated Aggression." *Journal of Personality and Social Psychology* 40 (4) (April): 687–700.

Blanchard, D. Caroline, and Robert J. Blanchard. 1984. "Affect and Aggression: An Animal Model Applied to Human Behavior." In *Advances in the Study of Aggression*, edited by Robert J Blanchard and D Caroline Blanchard, 1:1–62.

———. 1988. "Ethoexperimental Approaches to the Biology of Emotion." *Annual Review of Psychology*.

Blanchard, D.Caroline, and Robert J Blanchard. 2003. "What Can Animal Aggression Research Tell Us about Human Aggression?" *Hormones and Behavior* 44 (3) (September): 171–177. doi:10.1016/S0018-506X(03)00133-8.

Brigandt, Ingo. 2009. "Natural Kinds in Evolution and Systematics: Metaphysical and Epistemological Considerations." *Acta Biotheoretica* 57 (1-2) (July): 77–97. doi:10.1007/s10441-008-9056-7.

Brown, Rachael L. 2013. "Identifying Behavioral Novelty." *Biological Theory* (December 5). doi:10.1007/s13752-013-0150-y.

Carver, Charles S, and Eddie Harmon-jones. 2009. “Anger Is an Approach-Related Affect : Evidence and Implications.” *Psychological Bulletin* 135 (2): 183–204. doi:10.1037/a0013965.

Chevalier-Skolnikoff, S. 1974. “The Ontogeny of Communication in the Stumptail Macaque (Macaca Arctoides).”

Chevalier-Skolnikoff, S. 1973. “Facial Expression of Emotion in Nonhuman Primates.” In *... and Facial Expression: A Century of ...*, 11–89.

Defensor, Erwin B, Michael J Corley, Robert J Blanchard, and D Caroline Blanchard. 2012. “Facial Expressions of Mice in Aggressive and Fearful Contexts.” *Physiology & Behavior* 107 (5) (December 5): 680–5. doi:10.1016/j.physbeh.2012.03.024.

Delgado, Jose M. R. 1968. “Offensive-Defensive Behaviour in Free Monkeys and Chimpanzees Induced by Radio Stimulation of the Brain.” In *Aggressive Behavior*, edited by S. Garattini and E. B. Sigg, 109–119. New York: John Wiley and Sons.

Ekman, Paul. 1999. “Basic Emotions.” In *The Handbook of Cognition and Emotion*, edited by Tim Dalgleish and Mick Power, 45–60. Sussex, UK: John Wiley & Sons, Ltd.

Ekman, Paul, and WV Friesen. 1971. “Constants across Cultures in the Face and Emotion.” *Journal of Personality and Social ...*.

Ereshefsky, Marc. 2007. “Psychological Categories as Homologies: Lessons from Ethology.” *Biology & Philosophy* 22 (5) (September 29): 659–674. doi:10.1007/s10539-007-9091-9.

Frijda, Nico H. 1986. *The Emotions*. Edited by Knight Dunlap. *The Emotions*. Vol. 1. Studies in Emotion and Social Interaction. Cambridge University Press. doi:10.1093/0199253048.001.0001.

Frijda, Nico H., Peter Kuipers, and Elisabeth ter Schure. 1989. “Relations among Emotion, Appraisal, and Emotional Action Readiness.” *Journal of Personality and Social Psychology* 57 (2): 212–228. doi:10.1037//0022-3514.57.2.212.

Griffiths, Paul E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. Vol. 1997. University of Chicago Press.

———. 2007. “The Phenomena of Homology.” *Biology & Philosophy* 22 (5) (October 10): 643–658. doi:10.1007/s10539-007-9090-x.

Hitchcock, E, and V Cairns. 1973. “Amygdalotomy.” *Postgraduate Medical Journal* 49 (578) (December): 894–904.

Kruk, M R. 1991. "Ethology and Pharmacology of Hypothalamic Aggression in the Rat." *Neuroscience and Biobehavioral Reviews* 15 (4) (January): 527–38.

Leyhausen, Paul. 1979. *Cat Behavior: The Predatory and Social Behavior of Domestic and Wild Cats*. Translated by Batrbara A. Tonkin. 1St Editio. Taylor & Francis / Garland STPM Press.

Lipp, H. P., and R. W. Hunsperger. 1978. "Threat, Attack and Flight Elicited by Electrical Stimulation of the Ventromedial Hypothalamus of the Marmoset Monkey." *Brain, Behavior and Evolution* 15: 260–293.

Moyer, KE. 1976. "The Psychobiology of Aggression."

Owen, R. 1846. *Lectures on the Comparative Anatomy and Physiology of the Vertebrate Animals*. Vol. 2. Printed for Longman, Brown, Green, and Longmans.

Panksepp, Jaak. 1971. "Aggression Elicited by Electrical Stimulation of the Hypothalamus in Albino Rats." *Physiology & Behavior* 6 (4) (April): 321–9. doi:10.1016/0031-9384(71)90163-6.

———. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. 1st ed. Oxford University Press, USA.

Panksepp, Jaak, and Lucy Biven. 2012. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. W. W. Norton & Company.

Panksepp, Jaak, and MR Margaret R Zellner. 2004. "Towards A Neurobiologically Based Unified Theory of Aggression." *REVUE INTERNATIONALE DE ...* 17 (2): 37–61.

Parr, LA, BM Waller, SJ Vick, and KA Bard. 2007. "Classifying Chimpanzee Facial Expressions Using Muscle Action." *Emotion (Washington, DC)* 7 (1): 172–181. doi:10.1037/1528-3542.7.1.172.Classifying.

Pedersen, William C, Thomas F Denson, R Justin Goss, Eduardo a Vasquez, Nicholas J Kelley, and Norman Miller. 2011. "The Impact of Rumination on Aggressive Thoughts, Feelings, Arousal, and Behaviour." *The British Journal of Social Psychology / the British Psychological Society* 50 (Pt 2) (June): 281–301. doi:10.1348/014466610X515696.

Putnam, Hilary. 1969. "Brains and Behaviour." In *Analysis*, edited by Ned Block, 30:1–19. Mind, Language and Reality. Blackwell.

Remane, Adolph. 1971. "Die Grundlagen Des Natürlichen Systems: Der Vergleichenden Anatomie Und Der Phylogenetik."

Rieppel, Olivier. 2005. "Modules, Kinds, and Homology." *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 304 (1) (January 15): 18–27. doi:10.1002/jez.b.21025.

Roberts, W. W., M. L. Steinberg, and L. W. Means. 1967. "Hypothalamic Mechanisms for Sexual, Aggressive and Other Motivational Behaviors in the Oppossum Didelphis Virginiana." *Journal of Comparative Physiology and Psychology* 64: 1–15.

Sell, Aaron, John Tooby, and Leda Cosmides. 2009. "Formidability and the Logic of Human Anger." *Proceedings of the National Academy of Sciences of the United States of America* 106 (35) (September 1): 15073–8. doi:10.1073/pnas.0904312106.

Shaikh, M B, a Steinberg, and a Siegel. 1993. "Evidence That Substance P Is Utilized in Medial Amygdaloid Facilitation of Defensive Rage Behavior in the Cat." *Brain Research* 625 (2) (October 22): 283–94.

Siegel, Allan. 2004. *Neurobiology of Aggression and Rage*. 1st ed. Informa Healthcare.

Siegel, Allan, and Jeff Victoroff. 2009. "Understanding Human Aggression: New Insights from Neuroscience." *International Journal of Law and Psychiatry* 32 (4): 209–15. doi:10.1016/j.ijlp.2009.06.001.

Wenzel, JW. 1992. "Behavioral Homology and Phylogeny." *Annual Review of Ecology and Systematics* 23 (1992): 361–381.

Woodworth, C H. 1971. "Attack Elicited in Rats by Electrical Stimulation of the Lateral Hypothalamus." *Physiology & Behavior* 6 (4) (April): 345–53.

Are Systems Neuroscience Explanations Mechanistic?

Carlos Zednik

czednik@uos.de
Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany

Paper to be presented at: *Philosophy of Science Association 24th Biennial Meeting (Chicago, IL)*, November 2014

Abstract

Whereas most branches of neuroscience are thought to provide mechanistic explanations, systems neuroscience is not. Two reasons are traditionally cited in support of this conclusion. First, systems neuroscientists rarely, if ever, rely on the dual strategies of decomposition and localization. Second, they typically emphasize organizational properties over the properties of individual components. In this paper, I argue that neither reason is conclusive: researchers might rely on alternative strategies for mechanism discovery, and focusing on organization is often appropriate and consistent with the norms of mechanistic explanation. Thus, many explanations in systems neuroscience can also be viewed as mechanistic explanations.

1. Introduction

There is a widespread consensus in philosophy of science that neuroscientists provide *mechanistic explanations*. That is, they seek the discovery and description of the mechanisms responsible for the behavioral and neurological phenomena being explained. This consensus is supported by a growing philosophical literature on past and present examples from various branches of neuroscience, including molecular (Craver 2007; Machamer, Darden, and Craver 2000), cognitive (Bechtel 2008; Kaplan and Craver 2011), and computational neuroscience (Kaplan 2011). In contrast, one area that has received relatively little philosophical attention is *systems neuroscience*: the study of networks at various levels of brain organization. Do systems neuroscientists, like their colleagues in other branches of the discipline, seek the discovery and description of mechanisms? Answering this question is important for gaining an improved understanding of this exciting and increasingly influential area of research, but also for determining whether the various branches of neuroscience are unified by a common set of epistemic practices and explanatory norms.

Three research traditions can be distinguished within contemporary systems neuroscience. The first seeks the identification and description of networks at various levels of brain organization. The second seeks to reproduce the brain's behavioral dynamics and information-processing capacities through artificial neural network simulations. The third tradition specializes in the development of concise mathematical descriptions of the holistic behavior of biological as well as artificial brain networks. Several philosophical commentators have recently denied that systems neuroscientists in either one of these research traditions

provide mechanistic explanations. To a large extent, this denial is motivated by the observation that systems neuroscientists rarely invoke the heuristic strategies of *decomposition* and *localization* that are traditionally associated with mechanistic explanation (Bechtel and Richardson 1993; Silberstein and Chemero 2012; Varela, Thompson, and Rosch 2001). Moreover, the fact that systems neuroscientists often emphasize brain networks' global organization rather than their detailed composition is often considered to be evidence that these researchers have abandoned the mechanistic approach (Silberstein and Chemero 2012). If these commentators are correct, systems neuroscientists are quite unlike their colleagues in other branches of the discipline: they do not provide mechanistic explanations.

In what follows, I briefly introduce the three main research traditions within systems neuroscience (Section 2), and present the main reasons for thinking that researchers working within these traditions have abandoned mechanistic explanation (Section 3). Subsequently, I argue that these reasons are inconclusive (Section 4). For one, systems neuroscientists who eschew the strategies of decomposition and localization may appeal to alternative strategies for discovering mechanisms. For another, focusing on mechanistic organization rather than composition should not be viewed as a departure from mechanistic explanation. Indeed, such a focus is appropriate when a mechanism's organization is the principal determinant of that mechanism's behavior. Insofar as the discipline of systems neuroscience specializes in describing the organization of large network mechanisms in the brain, and insofar as the concept of organization remains poorly understood in philosophy of science, philosophers have much to gain from an improved understanding of mechanistic explanation in this increasingly important branch of neuroscientific research.

2. Three Research Traditions

Systems neuroscience is motivated by the observation that the brain is a complex system of networks of different kinds and at different levels of organization, as well as the observation that these networks are involved in the production of a wide range of behavioral and neurological phenomena. Within systems neuroscience, three conceptually distinct but mutually beneficial research traditions can be distinguished. Whereas the *network tradition* concerns the identification, description and topological analysis of brain networks, the *simulation tradition* specializes in the design of artificial neural network models to simulate the biological brain's behavioral dynamics and information-processing capacities. Both of these traditions are linked to the *complexity tradition*, the aim of which is to develop concise mathematical descriptions of the behavioral dynamics or information-processing capacities of artificial as well as biological brain networks.

The network tradition in systems neuroscience aims to identify and describe brain networks of different kinds and at different levels of organization. Thus, networks have been identified at the level of individual neurons within a population, at the level of neural populations (e.g. cortical columns) within a cortical region, and at the level of the brain as whole: networks of cortical regions. At each one of these levels, researchers in the network tradition must first determine how to identify a network's elements. Sometimes this choice is principled, as when biological principles are used to individuate nerve cells or cortical regions. At other times the choice is highly pragmatic, as when the network elements are chosen to

correspond with voxels in fMRI data or with the placement of electrodes in electrophysiological studies. After having identified a particular set of elements in this way, researchers in the network tradition must choose which kind of connections to pay attention to. Thus, networks might be defined over anatomical links such as synapses, but might also be defined over causal or functional links, typically operationalized as correlated activity over time. Once these choices have been made, brain networks can be identified via a variety of mapping and imaging techniques. These range from invasive methods such as histological studies of nerve cells and synaptic connections, to non-invasive methods such as structural, functional and diffusion MRI, among others (for a detailed overview of these methods, and of the network tradition in systems neuroscience as a whole, see: Sporns 2011).

Once brain networks have been identified, researchers in the network tradition typically represent and study these networks by invoking the mathematical framework of *graph theory*. A graph theoretical representation of a brain network characterizes the network's elements as nodes in a graph, and the anatomical, functional or causal links between these elements as (possibly weighted and/or directional) connections between individual nodes. Such graph theoretical representations facilitate the study of a network's global and local organization or *topology*. Thus, large-scale brain networks of all kinds have been shown to exhibit *small-world* topologies, that is, a high-degree of local clustering with short average path-lengths (Bassett and Bullmore 2006). Similarly, researchers have demonstrated the existence of *hub nodes* of relatively high degree, network *motifs* (small sub-graphs that are repeated throughout a network), and *modules*, i.e. densely interconnected communities of nodes with relatively sparse links to nodes in other communities (Sporns, Honey, and Kötter 2007; Sporns and Kötter 2004;

Meunier, Lambiotte, and Bullmore 2010). To date, these and other graph theoretic concepts have been used to understand the topology of brain networks in *C. elegans* (White et al. 1986; Izquierdo and Beer 2013), as well as cats and macaque monkeys (Sporns and Kötter 2004), and are likely to prove instrumental in the eventual success of the *Human Connectome Project* (Sporns, Tononi, and Kötter 2005) and other human brain mapping initiatives.

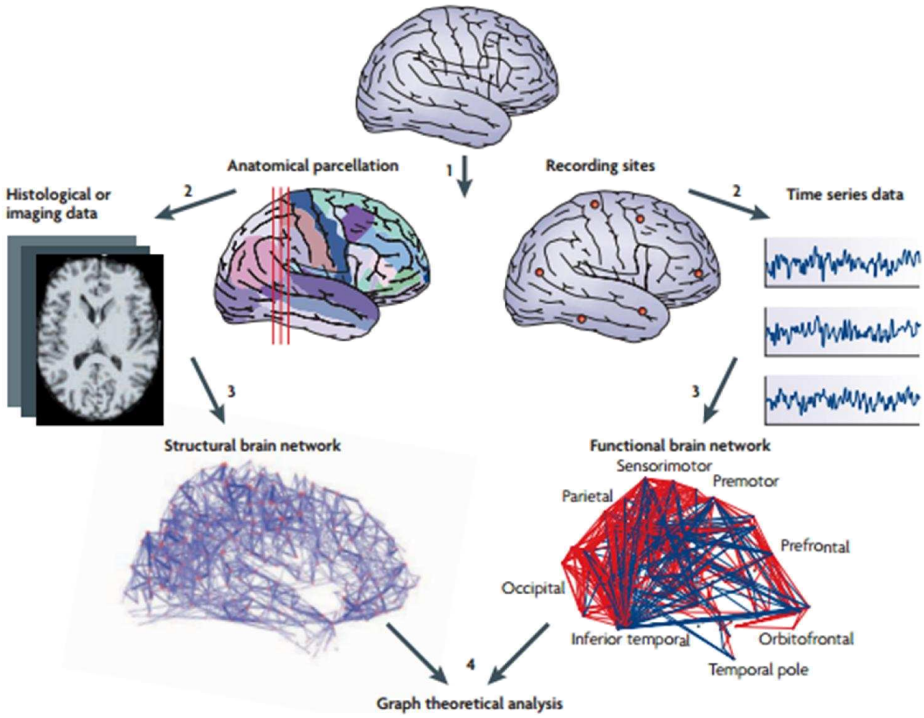


Figure 1. A schematic representation of model-development in the network tradition. The left pathway represents the development of structural network models; the right pathway represents the development of functional network models. Reprinted from Bullmore & Sporns (2009).

Whereas the network tradition seeks the description and topological analysis of brain networks, the simulation tradition aims to develop artificial neural network models to

reproduce the brain's behavioral dynamics or information-processing capacities. Thus, the primary purpose of these models is to understand the parameters under which brain networks might produce oscillations, synchronized firing patterns, and robustness to perturbation, as well as to understand when such networks are particularly good or bad at processing and integrating information. Although the relevant parameters are often local—determining e.g. the way in which individual network elements transform inputs to outputs—perhaps the most interesting research in this tradition seeks to understand the influence of topological parameters. Thus for example, Perez et al. (2011) have explored the extent to which random, small-world and scale-free networks exhibit local and global patterns of synchronization. Similarly, Tononi & Sporns (2003) have studied the relationship between the degree of modularity in a network and the degree of information integration—the ease by which information can be transmitted between any two network elements. In general, whereas the network tradition in systems neuroscience seeks to describe the kinds of networks that actually exist in the brain, the simulation tradition seeks to understand what these kinds of networks can do.

The third main research tradition within the discipline of systems neuroscience can be termed the *complexity tradition*. Building on decades of research in the discipline of complexity science, this tradition within systems neuroscience aims to develop concise mathematical models of behavior and information-processing in biological and artificial networks. These models are typically developed using the mathematical concepts and methods of *dynamical systems theory* and *information theory*. Whereas the former can be used to concisely characterize a network's behavior over time, the latter shows how information is processed and

integrated. One of the key features of these models is their relative simplicity. Whereas a particular network may consist of myriad reciprocally and non-linearly connected elements, its behavioral dynamics is often quite simple, exhibiting periodic oscillation or stability over time, and its information-processing efficient. The complexity tradition in systems neuroscience seeks to mathematically characterize these simple behaviors, which are often also characterized as the "emergent" properties of brain networks.

Although most research projects in systems neuroscience can be associated with either one of these three research traditions, it is not uncommon to see researchers combine the methods and results of two or more of them. Thus, researchers working in the simulation tradition increasingly seek to replace highly unrealistic artificial neural network models with "biologically inspired" network models that are rooted in the findings of the network tradition. Similarly, although many of the mathematical models developed in the complexity tradition were originally developed to characterize the global behavior of artificial neural networks, some of these models are proving themselves to be exceedingly useful for understanding the behavioral dynamics or information-processing of the biological brain. Notably, these combined research efforts are likely to be particularly influential in the future, since they apply the analytic power of computer simulations and mathematical analyses to increasingly accurate and biologically plausible models of the brain.

3. Abandoning Mechanistic Explanation

To what extent do researchers working within these three research traditions provide mechanistic explanations of behavioral and neurological phenomena? Mechanistic explanations center on descriptions or models of the mechanisms responsible for the phenomena being explained. Although there have been many different formulations of what constitutes a mechanism, the basic idea is that of an organized system of parts (or "entities") and operations (or "activities") whose changing properties over time exhibit a phenomenon of explanatory interest (Bechtel and Richardson 1993; Machamer, Darden, and Craver 2000; Bechtel and Abrahamsen 2010; Craver 2007). Notably, although most early contributions to the literature emphasize the characteristically diagrammatic nature of mechanistic explanations, many recent treatments acknowledge the possibility and prevalence of mathematical mechanism-descriptions. Thus, Kaplan & Craver (2011) have recently proposed a *model-mechanism-mapping constraint* (3M) intended to capture the requirements all mechanistic models must satisfy for them to be used in mechanistic explanations:

> "3M: In successful explanatory models…(*a*) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (*b*) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism." (Kaplan and Craver 2011, 611)

Before a mechanism can be described, it must be discovered: Researchers in neuroscience do not typically know in advance what the component parts and operations of a brain mechanism might be. According to Bechtel & Richardson's (1993) celebrated account of mechanism discovery, neuroscientists typically rely on the heuristic strategies of *decomposition* and *localization* to identify a mechanism's component parts and operations, and thus, to

develop a description or model of that mechanism. That is, they analyze the target phenomenon into a series or complex of relatively simple operations, break apart the physical system from which that phenomenon arises into a collection of smaller parts, and then study in detail the behavior of individual parts in order to link those parts to particular operations. Because of their perceived importance to mechanistic explanation in several disciplines, decomposition and localization are now often considered to be "the *sine qua non* of mechanistic explanation" (Silberstein and Chemero 2012, 3).

It is generally agreed that brain networks are mechanisms in the above sense. For one, they can clearly be viewed as organized systems of parts and operations. For another, insofar as many brain networks can be associated with particular behavioral or neurological processes, they can be said to exhibit various phenomena of explanatory interest. But although brain networks are mechanisms, some commentators have questioned whether the researchers who study these networks seek mechanistic explanations, as opposed to other kinds of explanations. Their questioning usually centers on the models being developed in systems neuroscience: what kinds of models these are, what kinds of properties they describe, and how these models are developed.

Unlike their colleagues in other branches of neuroscience, systems neuroscientists rarely invoke the heuristic strategies of decomposition and localization. This can be observed in all three of the research traditions introduced above. In the complexity tradition, researchers do not attempt to describe a brain network's component parts and operations at all, but just seek to describe its overall behavior. Thus, the mathematical models being developed in this

tradition are often *phenomenological models* rather than mechanistic models: they accurately describe the phenomenon being explained, without representing the causal structures and processes responsible therefore (Mauk 2000; Weiskopf 2011).

In the simulation tradition, the brain's behavioral dynamics and information-processing capacities are reproduced by way of artificial neural network models. Although these models can be construed as mechanistic models because they describe a network's component elements and connections, they are not usually developed via the heuristic strategies of decomposition and localization. In this tradition, the phenomenon being modeled is rarely analyzed into a series of localizable operations, and it is rare to see systems neuroscientists characterize the contributions of individual parts to the network mechanism's overall behavior. Thus, in Bechtel & Richardson's assessment, the simulation tradition

> "emphasizes systems whose dynamic behavior corresponds to the activity we want to explain, but in which the components of the system do not perform recognizable subtasks of the overall task…The overall architecture of the system—and especially the way components are connected—is what explains cognitive capacities, and not the specific tasks performed by the components. We have abandoned decomposition and localization." (Bechtel and Richardson 1993, 222–223)

Thus, rather than describe the contributions of a network's individual elements and connections by deploying the strategies of decomposition and localization, researchers in the simulation tradition focus on the contribution of the network's organization to its overall behavior.

Researchers working in the network tradition within systems neuroscience similarly focus on overall organization rather than individual components. Although developing a

network model of a biological brain network involves physically decomposing the brain or brain area being studied to determine the individual elements of the model, researchers rarely seek to describe in detail the structural features of these elements. Indeed, more often than not such network elements correspond to spatially-defined voxels or segments, rather than biologically-defined cortical structures that might plausibly be viewed as a mechanism's "working parts" (Craver 2007). In addition, it can be hard to interpret the connections between the elements of such a network model as a mechanism's component operations. Although some studies have begun to identify causal interactions between brain structures, far more is known about the brain's structural and statistical connectivity. Thus, although it may be known that a brain network's elements are structurally linked and/or statistically interrelated, it remains generally unclear exactly how these individual elements interact, and thus, unclear how they each contribute to the behavioral or neurological phenomena of explanatory interest.

In each one of the three research traditions, the focus on overall behavior and organization rather than detailed composition is often by necessity, rather than by choice. It has long been known (and current research in the network tradition has confirmed) that brain networks at all levels of organization often feature massively reciprocal and non-linear interactions (Varela, Thompson, and Rosch 2001; Sporns 2011). In such systems, the behavior of any individual component is at all times influencing, but also being influenced by, the behavior of the rest of the system. Thus, although it may be possible to physically individuate neurons, neural populations, and cortical regions, it can be difficult (or indeed, computationally intractable) to describe their individual contribution to the behavior of the system as a whole. What researchers in the simulation and complexity traditions of systems neuroscience have

shown, in contrast, is that a network's behavior can often be more usefully predicted and understood by focusing on holistic organizational properties instead. Thus, in Silberstein & Chemero's words,

> "rather than viewing the neurons, cell groups or brain regions as the basic unit of explanation, it is brain multiscale networks and their large-scale, distributed and non-local connections or interactions that are the basic unit of explanation." (Silberstein and Chemero 2012, 5)

To summarize: Unlike their colleagues in other branches of neuroscience, systems neuroscientists do not invoke the heuristic strategies of decomposition and localization to discover and describe mechanisms, and indeed, are frequently prevented from doing so due to the characteristic complexity of brain networks. Because their focus is placed squarely on network organization rather than detailed composition, there are good reasons to be skeptical about systems neuroscientists' commitment to mechanistic explanation.

4. Mechanistic Explanation in Systems Neuroscience

The previous section described reasons for believing that systems neuroscientists have abandoned mechanistic explanation. This section argues that these reasons rely on an unnecessarily narrow conception of mechanistic explanation. Mechanistic explanation need not invoke the heuristic strategies of decomposition and localization, and need not be focused on component parts and operations. Indeed, when a particular mechanism's behavior is largely determined by its overall organization, emphasizing this organization is appropriate and consistent with the norms of mechanistic explanation.

Consider first the suggestion that systems neuroscientists have abandoned mechanistic explanation just because they eschew the heuristic strategies of decomposition and localization. Taken at face value, this suggestion concerns the epistemic practices researchers invoke during the process of model development. Thus construed, however, it is wrong to assume that decomposition and localization are "the *sine qua non* of mechanistic explanation" as Silberstein & Chemero suggest. According to the conception of scientific discovery embraced by Bechtel & Richardson (1993), decomposition and localization greatly facilitate mechanism discovery by allowing researchers to quickly (albeit fallibly) traverse a conceptual space of "how-possibly" mechanistic models, with the goal of eventually identifying a "how-actually" model of the mechanism. However, decomposition and localization are not the only heuristic strategies that can be used for this purpose: there are a wide range of strategies to choose from.

Zednik (in press) has recently explored some alternative strategies for mechanism discovery. To cite just one example, consider the way researchers in evolutionary robotics invoke evolutionary algorithms to develop simulated mechanisms for *minimally cognitive* tasks (Harvey et al. 2005; Beer 2003). Because the simulated mechanisms that emerge from such evolutionary algorithms are relatively unconstrained by the ingenuity and design preferences of human researchers, many of them are characterized by features often seen as obstacles to mechanistic explanation: reciprocal non-linear interactions, a close integration of brain, body, and environment, and high sensitivity to temporal detail. Although it remains to be seen to what extent the discovery of such mechanisms in simulation can be used to make concrete inferences about analogous real-world mechanisms (for discussion see Webb (2009) and

responses to this target article), the evolution and analysis of these simulated mechanisms can surely be viewed as a heuristic strategy that helps to identify areas of the space of possible mechanistic models that merit further exploration. Thus, the approach adopted by evolutionary roboticists can be understood as a proof of concept that there exist heuristic strategies for mechanism discovery beyond decomposition and localization.

Considered as heuristic strategies, therefore, decomposition and localization are not in fact essential for mechanistic explanation. As a consequence, the mere fact that systems neuroscientists rarely invoke these heuristic strategies does not by itself show that they have abandoned mechanistic explanation. That said, an alternative way of understanding the claim that decomposition and localization are "the *sine qua non* of mechanistic explanation" concerns not the heuristic strategies themselves, but the result of applying these strategies: descriptions of the component parts and operations of mechanisms. Are such componential descriptions, as opposed to descriptions that mainly or exclusively describe a mechanism's organization, essential for mechanistic explanation?

Recall that mechanisms are organized systems of parts and operations that exhibit a particular phenomenon of explanatory interest, and that mechanistic explanations center on descriptions or models of such mechanisms. Moreover, recall that according to Kaplan & Craver's 3M constraint, models that figure in mechanistic explanations contain variables that "correspond to components, activities, properties, and organizational features of the target mechanism", the dependencies between which "correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism" (Kaplan and Craver 2011,

611). Notably, the second part of the 3M constraint is indicative of one of the most important norms of mechanistic explanation, viz. that such explanations should demonstrate how a target phenomenon "is situated in the causal structure of the world" (Craver 2013, 135). Exactly what such a demonstration would amount to is of course controversial, but one common measure of success is that a mechanistic model should render the represented mechanism amenable to interventions of manipulation and control (Woodward 2003; Craver 2007). That is, the model should represent (just) those properties of the mechanism that, when lesioned, activated, or otherwise influenced, effect predictable changes in the mechanism's behavior. Notably, although this norm is usually understood in terms of interventions on individual component parts or operations, it is also satisfied by models that facilitate interventions on the mechanism's overall organization.

Recall that many of the artificial and biological networks studied by systems neuroscientists consist of a large number of elements with reciprocal non-linear interactions. In these networks, organizational properties are frequently the primary determinants of overall behavior (Sporns 2011; Varela, Thompson, and Rosch 2001). Indeed, these properties can often be thought of as *order parameters*, the modification of which predicts changes in the network's overall behavioral dynamics and information processing capacities. Thus for example, Tononi & Sporns (2003) measure the effect varying degrees of modularity have on the degree of information integration within a network, presenting their results as a continuous function in which information integration is highest for networks with intermediate degrees of modularity, and lower for completely modular and completely homogeneous networks. Such studies reveal that network models can emphasize organizational properties instead of properties of

individual component parts and operations, while still rendering the relevant network mechanism amenable to interventions of manipulation and control. Insofar as this is one of the principal norms of mechanistic explanation, models that describe brain networks in ways that satisfy this norm should be considered mechanistic. Therefore, systems neuroscience explanations that center on such models can after all be viewed as mechanistic explanations.

How far can this line of reasoning be pushed? Although the network models being developed in the network and simulation traditions emphasize topological properties, they still provide some (admittedly not very detailed) insight into the individual component parts and operations of network mechanisms. Therefore, these models satisfy both parts of the 3M constraint, while additionally facilitating interventions of manipulation and control. But now compare these models to the models developed in the complexity tradition, the aim of which is to provide concise mathematical descriptions of brain networks' behavioral dynamics and information-processing capacities. As was discussed previously, such descriptions often amount to phenomenological models, i.e. models that describe the phenomena being explained without representing the causal structures and processes—and a fortiori, the mechanisms—responsible therefore. But not all models in the complexity tradition are pure phenomenological models. Indeed, some of the most interesting models in this tradition link variables that describe a network's global behavior to parameters that represent its topology. Although these models do not represent the relevant network mechanism's individual parts and operations, they do represent its organization in a way that renders it amenable for interventions of manipulation and control. Are these models mechanistic as well?

Whereas the network models developed in the network and simulation traditions are quite clearly mechanistic, and the systems neuroscientists in the network and simulation traditions can therefore be thought to provide mechanistic explanations, it is not so clear what to make of non-phenomenological models in the complexity tradition. On one hand, these models fail to describe what is often viewed as essential for mechanistic explanation: the relevant mechanism's component parts and operations. On the other hand, they do at least describe the mechanism's organization, and researchers often exploit this fact for the purposes of manipulation and control. Therefore, whether or not these models are mechanistic models depends on how these competing considerations are weighed against one another. Perhaps the following additional consideration can be used to break the tie in favor of the latter: Insofar as mechanistic models should *just* represent those properties that significantly contribute to a mechanism's behavior, and insofar as in brain networks these properties are often predominantly organizational, it seems that even those models in the complexity tradition that exclude compositional details entirely should be construed mechanistically.

5. Conclusion: Toward an Account of Mechanistic Organization

The preceding sections aimed to show that systems neuroscience—spanning the network, simulation, and complexity traditions—can often be thought to provide mechanistic explanations of behavioral and neurological phenomena. Although systems neuroscientists only rarely invoke the heuristic strategies of decomposition and localization, these strategies are not essential for mechanistic explanation. Moreover, and most importantly, although systems

neuroscientists typically emphasize organizational properties over and above the properties of individual component parts and operations, these properties can often be used for the purposes of manipulation and control. Thus, although they might pay far less attention to individual component parts and operations, systems neuroscientists seem to embrace many of the same explanatory methods and norms as their colleagues in other branches of neuroscience.

In closing, it is worth reflecting on the further philosophical significance of this mechanistic construal of systems neuroscience. Because systems neuroscientists specialize in characterizing the organization of a particular family of mechanisms—brain networks—their work is likely to be particularly beneficial to philosophers of science seeking an improved understanding of mechanistic organization. Although the concept of organization has long been included in philosophical definitions of "mechanism", it remains poorly understood, and indeed, its importance underestimated. Thus for example, one particularly influential treatment of mechanistic explanation presupposes that mechanisms are organized linearly, from "set up to termination conditions" (Machamer, Darden, and Craver 2000). In contrast, Bechtel & Abrahamsen (2010) have demonstrated that many mechanisms in neuroscience and biology are organized cyclically, and Levy & Bechtel (2013) have sought to explore the role of *motif*-like organizational building blocks in mechanisms in the life sciences. Insofar as systems neuroscientists have revealed many canonical forms of mechanistic organization in the brain—as well as many ways of studying this organization—it stands to reason that philosophers of science have much to gain from paying increased attention to this exciting and increasingly influential area of research.

References

Bassett, Danielle Smith, and Edward T. Bullmore. 2006. "Small-World Brain Networks." *The Neuroscientist* 12 (6) (December): 512–523.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, William, and Adele Abrahamsen. 2010. "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science." *Studies in History and Philosophy of Science Part A* 41 (3) (September): 321–333.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Beer, Randall D. 2003. "The Dynamics of Active Categorical Perception in an Evolved Model Agent." *Adaptive Behavior* 11 (4) (December 1): 209–243; discussion 244–305.

Bullmore, Edward T., and Olaf Sporns. 2009. "Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems." *Neuroscience* 10 (March): 186–198.

Craver, Carl F. 2007. *Explaining the Brain*. Oxford: Oxford University Press.

———. 2013. "Functions and Mechanisms: A Perspectivalist View." In *Functions: Selection and Mechanisms*, edited by Phillipe Huneman, 133–158. Dordrecht: Springer.

Harvey, Inman, Ezequiel A. di Paolo, Elio Tuci, Rachel Wood, and Matt Quinn. 2005. "Evolutionary Robotics: A New Scientific Tool for Studying Cognition." *Artificial Life* 11: 79–98.

Izquierdo, Eduardo J., and Randall D. Beer. 2013. "Connecting a Connectome to Behavior: An Ensemble of Neuroanatomical Models of C. Elegans Klinotaxis." *PLoS Computational Biology* 9 (2).

Kaplan, David M. 2011. "Explanation and Description in Computational Neuroscience." *Synthese* 183: 339–373.

Kaplan, David M., and Carl F. Craver. 2011. “The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective.” *Philosophy of Science* 78 (October): 601–627.

Levy, Arnon, and William Bechtel. 2013. “Abstraction and the Organization of Mechanisms.” *Philosophy of Science* 80 (2) (April): 241–261.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67 (1) (April): 1–25.

Mauk, Michael D. 2000. “The Potential Effectiveness of Simulations versus Phenomenological Models.” *Nature Neuroscience* 3 (7) (July): 649–651.

Meunier, David, Renaud Lambiotte, and Edward T. Bullmore. 2010. “Modular and Hierarchically Modular Organization of Brain Networks.” *Frontiers in Neuroscience* 4 (December).

Perez, Toni, Guadalupe C. Garcia, Victor M. Eguiluz, Raul Vicente, Gordon Pipa, and Claudio Mirasso. 2011. “Effect of the Topology and Delayed Interactions in Neuronal Networks Synchronization.” *PLoS One* 6 (5).

Silberstein, Michael, and Anthony Chemero. 2012. “Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences.” In *Philosophy of Science Association 23rd Biennial Meeting*. San Diego, CA.

Sporns, Olaf. 2011. *Networks of the Brain*. Cambridge, MA: MIT Press.

Sporns, Olaf, Christopher J. Honey, and Rolf Kötter. 2007. “Identification and Classification of Hubs in Brain Networks.” *PLoS One* 2 (10).

Sporns, Olaf, and Rolf Kötter. 2004. “Motifs in Brain Networks.” *PLoS Biology* 2 (11) (November): e369.

Sporns, Olaf, Giulio Tononi, and Rolf Kötter. 2005. “The Human Connectome: A Structural Description of the Human Brain.” *PLoS Computational Biology* 1 (4) (September): e42.

Tononi, Giulio, and Olaf Sporns. 2003. “Measuring Information Integration.” *BMC Neuroscience* 4 (31) (December 2).

Varela, Francisco, Evan Thompson, and Eleanor Rosch. 2001. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Webb, Barbara. 2009. “Animals Versus Animats: Or Why Not Model the Real Iguana?” *Adaptive Behavior* 17 (4) (July 28): 269–286.

Weiskopf, Daniel A. 2011. “Models and Mechanisms of Psychological Explanation.” *Synthese*.

White, J. G., E. Southgate, J. N. Thomson, and S. Brenner. 1986. “The Structure of the Nervous System of the Nematode Caenorhabditis Elegans.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 314: 1–340.

Woodward, Jim. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Zednik, Carlos. in Press. “Heuristics, Descriptions, and the Scope of Mechanistic Explanation.” In *How Does Biology Explain? An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences*, edited by Christophe Malaterre and Piere-Alain Braillard. Dordrecht: Springer.

Likelihood and Consilience: On Forster's Counterexamples to the Likelihood Theory of Evidence

Jiji Zhang

Department of Philosophy, Lingnan University

Kun Zhang

Max Planck Institute for Intelligent Systems

Abstract

Forster presented some interesting examples having to do with distinguishing the direction of causal influence between two variables, which he argued are counterexamples to the likelihood theory of evidence (LTE). In this paper, we refute Forster's arguments by carefully examining one of the alleged counterexamples. We argue that the example is not convincing as it relies on dubious intuitions that likelihoodists have forcefully criticized. More importantly, we show that contrary to Forster's contention, the consilience-based methodology he favored is accountable within the framework of the LTE.

1 Introduction

Forster (2006) presented some putative counterexamples to what he called a/the likelihood theory of evidence (LTE):

> "The Likelihood Theory of Evidence (LTE): The observed data are relevant to the comparison of simple hypotheses (or models) only via the likelihoods of the simple hypotheses being compared (or the likelihood functions of the models under comparison)." (321)

The LTE entails that if the likelihood of one hypothesis relative to a given body of data — i.e., the probability of obtaining the data given the hypothesis — is the same as that of another hypothesis, then the hypotheses cannot be distinguished based on the data alone. Forster challenged this consequence with examples in which the data, he argued, favor one hypothesis over another even though the two have the same likelihood.

For those concerned with causal inference, Forster's examples are particularly interesting because they have to do with distinguishing the direction of causal influence between two random variables. Forster contended that his examples demonstrate a distinctive methodology based on Whewell's notion of "consilience of inductions" (Whewell 1858; Forster 1988), which cannot be captured by a likelihoodist or Bayesian philosophy of science that subscribes to the LTE.

Our purpose in this paper is twofold, one critical and one positive. First, in section 2, we argue that Forster's challenge to the LTE is based on denying a basic, well-argued thesis of likelihoodism. The thesis is that the evidential bearing of a body of data on a given statistical hypothesis is essentially relative, depending on the alternative against which the hypothesis is assessed. The apparent force of Forster's counterexamples, we argue, relies upon embracing an intuition that likelihoodists (e.g., Hacking 1965; Royall 1997; Sober 2008) have forcefully criticized — the intuition that a statistical hypothesis, taken alone, can be rejected or shown to be false by data. At best, therefore, Forster's argument begs an important question against the likelihoodist.

Second, and more importantly, we aim to vindicate Forster's preferred methodology using likelihoods. We show in section 3 that there is a systematic connection between likelihood and the kind of consilience Forster emphasized. Forster is right that considerations of consilience are evidentially relevant. However, such relevance, we contend, is

reflected in likelihoods.

Due to the space limit, we will focus on Forster's example featuring discrete variables, but our points extend straightforwardly to his example with continuous variables, as we will briefly comment in section 4.

2 On Forster's challenge to the LTE

For the likelihoodist, a thesis of fundamental importance is what Royall (1997) called the "relativity of evidence". A body of data constitutes evidence for or against a statistical hypothesis *only* relative to some alternative hypothesis. For example, getting ten heads straight in tossing a coin is not evidence against the coin being fair *simpliciter*. It disconfirms the fair-coin hypothesis in reference to some alternative hypothesis, e.g., the hypothesis that the coin is a trick coin with heads on both sides, or that the coin is so biased towards one side that the chance of landing head in each flip is 0.9. But when compared to certain other alternatives, e.g., the hypothesis that the coin is a trick coin with tails on both sides, the observations favor the fair-coin hypothesis. The evidential bearing of the data on the fair-coin hypothesis is thus relative to the alternative being considered; evidential statements are essentially contrastive in form.

Detailed and compelling arguments for this view were elegantly presented by, among others, Royall (1997, 65-68) and Sober (2008, 48-52), and we shall not repeat them here. Suffice it to say that objections to the likelihood account of evidence that rely on denying the relativity of evidence begs an important question. We will argue that Forster's challenge ends up question-begging in this way.

It is not obvious that Forster ran afoul of the relativity of evidence. His counterexamples apparently pit a hypothesis against another. Here is the first version of the example

we will focus on in this paper. Suppose that two variables X and Y are related by a simple law: $Y = X + U$, where X is a variable taking positive integer values, and U is an unobserved variable — error term — taking one of two values: 0.5 or -0.5, with equal probability. Suppose also that data are generated by twenty independent trials, with $X = 4$ in each trial. As it happens, in ten of the twenty trials, Y is observed to be equal to 3.5 (i.e., the values of U in those trials are -0.5), and in the other ten trials, Y is observed to be equal to 4.5 (i.e., the values of U in those trials are 0.5).

Let us use X_i, Y_i, etc. to model the ith trial. Consider now two hypotheses. One is the true hypothesis, which Forster referred to as Hypothesis A: $Y_i = X_i + U_i$ $(i = 1, ..., 20)$, and the error terms U_i's are independently and identically distributed (i.i.d.) such that $P(U_i = -0.5) = P(U_i = 0.5) = 1/2$.

The other hypothesis is referred to as Hypothesis B (for Backwards): $X_i = Y_i + V_i$ $(i = 1, ..., 20)$, and the error terms V_i's are i.i.d. such that $P(V_i = -0.5) = P(V_i = 0.5) = 1/2$.[1]

In the first version of the example, Forster considered these two hypotheses as such, and treated the exogenous variable in each hypothesis as non-random or given. Specifically, in A, X_i's are not treated as random variables, but Y_i's are (because U_i's are); in B, Y_i's are not treated as random variables, but X_i's are (because V_i's are). For these hypotheses, as Forster pointed out, only *conditional* likelihoods are well defined. For A, the conditional likelihood is the probability of obtaining the observed values of Y_i under hypothesis A, given the values of X_i, which is $(1/2)^{20}$; for B, the conditional likelihood is the probability of obtaining the observed values of X_i under hypothesis B, given the

[1]Forster used the same symbol U to denote the error terms in both hypotheses, which is potentially misleading. To avoid confusions, we use V to denote the error term postulated by the backwards hypothesis.

values of Y_i, which is also $(1/2)^{20}$.

According to Forster (2006),

> "The example is already a counterexample to LTE in the following sense: We are told that either A or B is true, and we can tell from the data that A is true and B is false. But there is nothing in the *likelihoods* that distinguishes between them." (328, original emphasis)

We will return to Forster's claim that one can tell from the data that A is true and B is false. For now let us focus on a more basic problem with this version of the example. The problem is that the two hypotheses concern *different* random variables: the random variable in A is Y (or more accurately, $\langle Y_1, ..., Y_{20}\rangle$), and the random variable in B is X (or $\langle X_1, ..., X_{20}\rangle$). However, a presupposition of LTE is that the hypotheses in question concern a *common* set of random variables: the hypotheses imply probability distributions over these variables, and the data are observations of their values. Royall, for example, made it explicit in his influential formulation of the law of likelihood:

> "If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$ ... ." (Royall 1997, 3)[2]

Clearly the present version of the example does not satisfy the presupposition. Thus, for likelihoodists like Royall, it does not make sense to talk about the evidential support of A versus B.

[2]Royall seemed to attribute this formulation to Hacking (1965), but as far as we can see, Hacking did not formulate his law of likelihood in precisely these terms.

To be fair, Forster was quick to acknowledge that a subscriber to LTE could easily respond to this version of the example by denying that LTE should apply to such "incomplete" hypotheses. He put the subscriber's complaint in the following terms:

> "They might insist that the example violates the principle of total evidence because the likelihoods are not relative to the full data, even though there are no data "hidden from view", or withheld in any way." (Forster 2006, 328)

This, in our view, is a misdiagnosis on behalf of the the friends of LTE. The principle of total evidence is about *what* evidence to take into account, but the LTE is about the evidential bearing of *given* evidence on the comparison of hypotheses. It is perfectly sensible to ask whether a certain *part* of the data supports one hypothesis against another (though one should take total evidence into account, if possible, when updating beliefs or judgements). In the present case, for example, there is no problem comparing, based on conditional likelihoods, hypothesis A with, say, A^*: $Y_i = X_i + U_i, i = 1, ..., 20$, and $P(U_i = -0.5) = 1/4$ and $P(U_i = 0.5) = 3/4$. A and A^* are as "incomplete" as A and B are, but they are about the same random variables, and hence are comparable given the data. By contrast, A and B as such are incomparable[3] because they concern entirely different random variables.

Why are hypotheses incomparable if they are about different random variables? This is connected to the thesis of evidential relativity. To see the matter clearly, it helps to consider a simpler case. Suppose two coins are each flipped independently for twenty times. Of the first coin, all of the twenty flips turn up heads; of the second coin, half of the flips turn up heads and half turn up tails. Consider two hypotheses: (1) the first

[3]By "incomparable" we mean only that the hypotheses are not subject to *evidential* comparison. They may still be comparable in terms of rational credences or someone's personal credences.

coin is fair, and (2) the second coin is fair. The observations on the first coin — call them D_1 — are irrelevant to hypothesis (2) (in the absence of any background knowledge or assumption linking the two coins). So it does not make sense to say that D_1 provide evidence for or against (1) versus (2). Similarly, we cannot say that D_2, the data on the second coin, provide evidence for or against (2) versus (1).

However, it may be tempting to think that the degree to which D_2 support (2) is greater than the degree to which D_1 support (1). After all, it seems intuitive that (1) fits D_1 very poorly while (2) fits D_2 rather well. If so, it would be fair to say that (1) and (2) are comparable after all, given the combined data $D = \langle D_1, D_2 \rangle$. But according to the relativity of evidence, there is no such thing as the degree to which D_2 support (2) *simpliciter* or that to which D_1 support (1) *simpliciter*. D_1 confirm or disconfirm (1) only in contrast to some other hypothesis concerning the outcomes of flipping coin 1, and D_2 confirm or disconfirm (2) only in contrast to some other hypothesis concerning the outcomes of flipping coin 2. Hence, it does not make sense to say that D_2 support (2) better than D_1 support (1).

Therefore, from the likelihoodist point of view, the basic problem with the present 'counterexample' is not so much a violation of the principle of total evidence as a juxtaposition of incomparable hypotheses, and the incomparability is closely related to the relativity of evidence. Forster's neglect of this point signals his denial of the relativity of evidence.

In any case, Forster did develop the example into one with comparable hypotheses. Treat both $X_i's$ and $Y_i's$ as random variables. To A add the assumption that X_i and U_i are statistically independent, and that $P(X_i = x_i) = 1$, where x_i is the observed value of X on the ith trial. To B add the assumption that Y_i and V_i are statistically independent,

and that $P(Y_i = y_i) = 1$, where y_i is the observed value of Y on the ith trial. That is, the marginal distributions are specified in an *ad hoc* way to the effect that whatever values the hypothesized exogenous variables actually take, the (constructed) hypotheses entail that they take those values with probability 1. Such marginals are objectionable and useless in practice for a number of reasons, but we will leave them aside. Following Forster, call the resulting hypotheses A' and B'. They have the same likelihood.[4]

$$\begin{aligned} L(A') &= \prod_i P_{A'}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{A'}(Y_i = y_i | X_i = x_i) P_{A'}(X_i = x_i) = (1/2)^{20} \\ L(B') &= \prod_i P_{B'}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{B'}(X_i = x_i | Y_i = y_i) P_{B'}(Y_i = y_i) = (1/2)^{20} \end{aligned}$$

According to Forster, this constitutes a counterexample to the LTE because despite the equality of likelihoods, one can tell from the data that B' is false and A' is true. Here is his argument. B' entails that V_i and Y_i are independent (for every i):

$$P(V_i = 0.5 | Y_i = 3.5) = P(V_i = 0.5 | Y_i = 4.5) = 1/2$$

or equivalently, $P(X_i = 4 | Y_i = 3.5) = P(X_i = 5 | Y_i = 4.5) = 1/2$. Call this consequence B'_1. However, from the data we see that the relative frequency of $X = 4$ in the trials in which $Y = 3.5$ is 1, and the relative frequency of $X = 5$ in the trials in which $Y = 4.5$ is 0. Hence the data show that B'_1 is false, and so B' is false.

Formulated this way, the argument is clearly not contrastive, and seems akin to the *probabilistic modus tollens* that has been resolutely refuted by likelihoodists (Sober 2008, 51-53). It is not impossible, just very improbable, to obtain the data as they are even if

[4]Throughout the paper, we use upper case letters to denote variables and the corresponding lower case letters to denote values of the variables.

B_1' is true. A more charitable reading is that Forster did not literally mean that B_1' is shown to be false, but that the data overwhelmingly disconfirm B_1' relative to Not-B_1'. However, Not-B_1' is a complex class of hypotheses. Relative to some members in the class, the data are evidence against B_1', but relative to others, e.g. that $P(X_i = 4|Y_i = 3.5) = 0 \neq P(X_i = 5|Y_i = 4.5) = 1$, the data are arguably evidence for B_1'. In the absence of a well-grounded prior over these members, it is hard to make sense of the sweeping claim that the data seriously disconfirm B'-1 in favor of its logical negation.

Therefore, if we take the relativity of evidence seriously, the right way to state Forster's intuition is that the data provide evidence against B'-1 in reference to the given alternative A'. More accurately, the data disconfirm B_1' relative to A_1': $P(X_i = 4|Y_i = 3.5) = 1 \neq P(X_i = 5|Y_i = 4.5) = 0$, which is entailed by A'. Indeed, the evidence against B_1' versus A_1' is overwhelming, judging either intuitively or by a formal measure such as likelihood ratio.

But the fact that the data constitute weighty evidence against B_1' versus A_1' does not entail that the data are weighty evidence against B' versus A'. A_1' and B_1' are just parts of what A' and B' have to say about the data at hand; they are about the conditional probability of X_i given Y_i. But A' and B' also have implications for the marginal probability of Y_i. A' entails A_2': $P(Y_i = 3.5) = P(Y_i = 4.5) = 1/2$ (for every i), whereas B' entails B_2': $P(Y_i = y_i) = 1$, where y_i is the actual observed value of Y_i. How do the data bear on A_2' versus B_2'?

Essentially the same question was addressed very early on in Royall (1997)'s elaborate defense of likelihoodism. Shortly after he described the law of likelihood, he considered and refuted a putative counterexample (13-15). Suppose an ordinary-looking deck of 52 cards is well-shuffled. We turn over the top card and find it to be the ace of diamonds.

According to the law of likelihood, the observation supports the hypothesis that it is a trick deck consisting of 52 aces of diamonds against the hypothesis that the deck is ordinary. This may sound counterintuitive; intuitively the trick-deck hypothesis is not rendered more probable or believable than the ordinary-deck hypothesis based on the observation. But the evidential judgment is perfectly consistent with the intuition, for the question of credence is different from that of evidence. Even though the observation supports the trick-deck hypothesis against the ordinary-deck hypothesis, the former, in ordinary circumstances, is much less credible prior to the observation and may well end up less credible overall despite the positive evidence.

By the same token, for every trial in Forster's example, the observation of $Y_i = y_i$ supports the hypothesis that $P(Y_i = y_i) = 1$ against the hypothesis that $P(Y_i = 3.5) = P(Y_i = 4.5) = 1/2$, and overall the data favor B_2' over A_2'. Again, this evidential judgment should not be conflated with the judgment that the data render B_2' more credible than A_2'. In normal circumstances, there are a number of reasons to regard B_2' as much less plausible than A_2', prior to considering the evidence, and the evidential support may well be insufficient to overcome the initial implausibility.

The upshot is that the data are evidence for A_1' versus B_1', but also constitute evidence against A_2' versus B_2'. There is no compelling reason to think that the data *alone* favor the conjunction of A_1' and A_2' over that of B_1' and B_2' (or the other way around). The LTE, we conclude, is not threatened by the example.

3 Likelihood and consilience

Forster's positive insight, however, is not to be ignored. As he put it, Hypothesis B' suffers from a lack of "consilience". Given B', the probability distribution of the error

term V can be measured or estimated under two conditions: when $Y = 3.5$ and when $Y = 4.5$, but the two estimates do not "jump together": the empirical distribution of V estimated from the group of $Y = 3.5$ is very different from that of V estimated from the group of $Y = 4.5$. In contrast, Hypothesis A' does not have this problem (though in this case it does not display interesting consilience due to the lack of variation in X). We agree with Forster that this kind of consilience or lack thereof is evidentially significant, but we submit that the contrast is reflected in the comparison of likelihoods.

To show this, it helps to consider yet another version of the example Forster discussed. This version specifies the two hypotheses in the standard and perhaps most natural way, which assumes that X_i's and Y_i's are i.i.d. Under the i.i.d assumption, the best fitting marginal of X is $P(X = 4) = 1$. Add this marginal of X, together with the i.i.d assumption and that of the independence between X and U, to A, and call the resulting hypothesis A''. Similarly, the best fitting marginal of Y is $P(Y = 3.5) = P(Y = 4.5) = 1/2$. Add this marginal of Y, together with the i.i.d assumption and that of the independence between Y and V, to B, and call the resulting hypothesis B''. In this example, A'' happens to be the same as A', so $L(A'') = (1/2)^{20}$. But B'' is different from B', and has a much lower likelihood:

$$\begin{aligned} L(B'') &= \prod_i P_{B''}(X_i = x_i, Y_i = y_i) \\ &= \prod_i P_{B''}(X_i = x_i | Y_i = y_i) P_{B''}(Y_i = y_i) = (1/2)^{40} \end{aligned}$$

The difference in likelihoods accords well with the intuition that the data favor A'' over B''. But there is a "mystery" according to Forster. The likelihoods seem to differ just because of the difference in the parts contributed by the added marginals, but why should that matter? Intuitively, "the generation of the independent or exogenous variable [is] an inessential part of the causal hypothesis."

We share the latter intuition. In particular, we are sympathetic with the view that a defining feature of a *causal* relationship is that the relationship remains invariant under suitable interventions of an exogenous cause that change its marginal distribution (Woodward 2003). But it does not follow that marginals are irrelevant in causal inference. They are especially relevant to the kind of problems under discussion: distinguishing the direction of causal influence. In the present case, for example, X is hypothesized as the cause in only one of the hypotheses; in the other hypothesis it is modeled as the effect. The marginal distribution of X is relevant to judging, for example, how well the other hypothesis, by treating X as endogenous, fits the observations on X, compared to the former hypothesis that treats X as exogenous.

The right explanation in our view of the difference between the likelihoods actually agrees nicely with Foster's consideration of consilience. Notice that the lack of consilience under B highlighted by Forster corresponds to the statistical dependence of V on Y as shown in the data. A convenient measure of statistical dependence between random variables is known as *mutual information* (Cover and Thomas 1991, 18). The mutual information between two random variables Z and W is defined as:

$$\mathbf{I}(Z,W) = \mathbb{E}\log\frac{P(Z,W)}{P(Z)P(W)} = \mathbb{E}(\log P(Z,W) - \log P(Z) - \log P(W))$$

where $\mathbb{E}(\cdot)$ denotes the expectation (with respect to $P(Z,W)$). The mutual information is a way to measure the difference between the joint distribution $P(Z,W)$ and the product of the marginals $P(Z)P(W)$, and hence a way to measure the statistical dependence between Z and W. It is non-negative and is equal to zero just in case Z and W are independent.

Given i.i.d samples from the joint distribution $P(Z,W)$, we can use the following sample approximation of the mutual information, by replacing the expectation with the

sample mean, and P with the corresponding empirical distribution $\hat{P}$ (where probabilities are estimated by sample frequencies):

$$\hat{\mathbf{I}}(Z,W) = \frac{1}{n}\sum_i(\log \hat{P}(z_i,w_i) - \log \hat{P}(z_i) - \log \hat{P}(w_i))$$

where n is the sample size. This provides a measure of dependence as shown in the samples. Accordingly, in Forster's example, $\hat{\mathbf{I}}(X,U)$ and $\hat{\mathbf{I}}(Y,V)$ can be regarded as plausible measures of the lack of consilience in A'' and B'', respectively.

Some calculations are in order. Given the data in Forster's example, the empirical joint distribution of X and Y puts half of the mass on $\langle X = 4, Y = 3.5\rangle$ and half of the mass on $\langle X = 4, Y = 4.5\rangle$. It follows that the empirical joint distribution of X and $U = Y - X$ puts half of the mass on $\langle X = 4, U = -0.5\rangle$ and half of the mass on $\langle X = 4, U = 0.5\rangle$; and the empirical joint distribution of Y and $V = X - Y$ puts half of the mass on $\langle Y = 3.5, U = -0.5\rangle$ and half of the mass on $\langle Y = 4.5, U = 0.5\rangle$. From these it is easy to calculate, taking 2 as the base of the logarithm to simplify the numbers, that $\hat{\mathbf{I}}(X,U) = 0$ and $\hat{\mathbf{I}}(Y,V) = 1$. These, we repeat, are plausible measures of the lack of consilience in Forster's sense.

Consider now the log-likelihoods of A'' and B'', taking again 2 as the base of the logarithm: $l(A'') = -20$ and $l(B'') = -40$. The difference between them per datum is $20/20 = 1$, which is precisely the difference between $\hat{\mathbf{I}}(Y,V)$ and $\hat{\mathbf{I}}(X,U)$.

This is not a numerical accident. The log-likelihood of A'' can be written as:

$$l(A'') = \sum_i \log P_{A''}(x_i,y_i) = \sum_i \log P_{A''}(x_i,u_i) = \sum_i(\log P_{A''}(x_i) + \log P_{A''}(u_i))$$

Because for each i, $\langle X_i = x_i, Y_i = y_i\rangle$ and $\langle X_i = x_i, U_i = u_i = y_i - x_i\rangle$ are descriptions of the same event. Since the marginals in A'' are specified as the corresponding empirical

distributions — $P_{A''}(X) = \hat{P}(X)$ and $P_{A''}(U) = \hat{P}(U)$ — we have

$$\begin{aligned} l(A'') &= \sum_i (\log \hat{P}(x_i) + \log \hat{P}(u_i)) \\ &= \sum_i \log \hat{P}(x_i, u_i) - \sum_i (\log \hat{P}(x_i, u_i) - \log \hat{P}(x_i) - \log \hat{P}(u_i)) \\ &= \sum_i \log \hat{P}(x_i, u_i) - n\hat{\mathbf{I}}(X, U) \end{aligned}$$

Similarly,

$$l(B'') = \sum_i \log \hat{P}(y_i, v_i) - n\hat{\mathbf{I}}(Y, V)$$

Note further that for every i, $\hat{P}(x_i, u_i) = \hat{P}(y_i, v_i) = \hat{P}(x_i, y_i)$. Hence,

$$l(A'') - l(B'') = n(\hat{\mathbf{I}}(Y, V) - \hat{\mathbf{I}}(X, U))$$

Therefore, there is here a systematic connection between the comparison of likelihoods and the comparison of how hypotheses fare in terms of "consilience of inductions" highlighted by Forster. The evidential significance of consilience, or at least one plausible interpretation of it, is not beyond the grip of the LTE.

4 Conclusion

Whether or not the LTE can survive other challenges, Forster's examples, we conclude, are not convincing counterexamples. We have only examined one of his examples in this paper, but the other example, set up in linear models with continuous variables, employs parallel devices and arguments, to which our points in section 2, *mutatis mutandis*, carry over. The apparent force of his examples hinges on an (implicit) denial of the basic tenet of likelihoodism, i.e., the thesis of the relativity of evidence. Since the denial is based on no argument but dubious intuitions that have been forcefully criticized by likelihoodists, Forster's criticism is at best question begging.

More interestingly, we showed a way to vindicate Forster's preferred consilience-based methodology within the framework of the LTE, by establishing a systematic connection between likelihood and (one plausible interpretation of) consilience. The connection holds much more generally for such causal models than we can show in this paper (Hyvärinen and Smith 2013; Zhang et al. forthcoming). In particular, Hyvärinen and Smith (2013, 115) presented a similar result on linear models, which is applicable to Forster's example with continuous variables. Whether similar connections exist in contexts other than this sort of causal inference problems is worth exploring.

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: John Wiley and Sons.

Forster, M. (1998). Unification, explanation, and the composition of causes in Newtonian mechanics. *Studies in the History and Philosophy of Science*, **19**: 55–101.

Forster, M. (2006). Counterexamples to a likelihood theory of evidence. *Minds and Machines*, **16**(3): 319–338.

Hacking, I. (1865). *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, **14**: 111–152.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, Fla.: Chapman and Hall.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Whewell, W. (1858). *Novum Organon Renovatum*. London: John W. Parker.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford and New York: Oxford University Press.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (forthcoming). On estimation of functional causal models: Post-nonlinear causal model as an example. *ACM Transactions on Intelligent Systems and Technologies*.