Seventh Quadrennial Fellows Conference of the Center for Philosophy of Science

12-14 June 2012; Mugla, Turkey

Version: July 28, 2012

PhilSci ARCHIVE

Table of Contents

CULTURE AND TRANSITIONS IN INDIVIDUALITY

Paulo C. Abrantes*

Several biologists and philosophers have been arguing, for a while now, that a Darwinian evolutionary dynamics might take place not only in the distribution of phenotypic traits in a particular kind of population, but also in the very dimensions that are used to track those, bringing about new kinds of populations, given certain special circumstances. These "major" evolutionary transitions have sometimes been described as transitions in *individuality*. In this depiction, natural selection (maybe combined with other causes) often brings about new kinds of individuals, whose evolutionary dynamics takes place in a novel way. This topic became a big concern since the groundbreaking works of Buss (1987), Maynard-Smith and Szathmáry (1997), and Michod (1999). Godfrey-Smith's 2009 book follows this trend by emphasizing that "evolutionary processes are themselves evolutionary products" (2009, 15). One of the chief thesis he puts forth, by pushing population thinking even further, is that a transition in individuality is fully accomplished when a new, "paradigmatic", Darwinian population emerges. In collective entities, where there are nested populations embodied in one individual, the higher and the lower level populations follow different evolutionary paths during a major transition: the latter ones usually change their Darwinian status from a "paradigmatic" to a "marginal" one. This process of "de-Darwinization" of the lower level populations - as Godfrey-Smith describes the evolutionary transition taking place at that level (Ibid., 100) -, can be tracked by significant changes in the values of a set of parameters that describe their evolutionary dynamics or "evolvability" (Ibid., 41). The process of de-Darwinization of the populations of cells that make up multicellular organisms is a well-known case. In this paper, I want to investigate

* Department of Philosophy and Institute of Biological Sciences, University of Brasília, Brazil.
pccabr@gmail.com
Published version: http://www.cfh.ufsc.br/~nel/rumos.html

whether it is fruitful to describe the role that culture begins to play at some point in the Hominin lineage - arguably that of the emergence of a new inheritance system on top of the genetic inheritance system and coevolving with it -, as being a transition in individuality.

(I) REPRESENTING DARWINIAN DYNAMICS

Godfrey-Smith criticizes, in his book, previous attempts to give an abstract "summary" of the essential elements that are required for describing evolution in Darwinian terms (2009, 17). His way to open a new trail in what he calls the "classical approach" is to start with a "minimal concept" of a Darwinian population - which just requires that there be variation in the traits of individuals in a population that affect their reproduction and that part of this variation be heritable.[1]

The 'minimal concept' - associated with a "kind of change", evolution by natural selection - is permissive and includes much more than the paradigmatic cases of Darwinian populations (Godfrey-Smith, 2011, 67). To avoid the pitfalls of those attempts in the classical tradition, he aims to describe not only the purportedly paradigmatic cases of Darwinian populations, but also go into the marginal cases, that don't have all the features of the former ones. The particular way a kind of population located in this spectrum evolves depends on further features that are not specified by the minimal concept, requiring new parameters to describe its dynamics. In other words, the minimal concept provides just a "set up" and has to be complemented with "middle-level" theories or models to take into account the diversity of living beings and, more generally, of systems whose dynamics can be fruitfully described in populational-Darwinian terms (Ibid., 39; cf. 31).

Starting with the minimal concept as a scaffolding, Godfrey-Smith proposes a "spatial" representation in which the chief features of Darwinian populations, concerning their evolvability, are quantified in order to tell paradigmatic from marginal cases. This

[1] Godfrey-Smith criticizes the "replicator approach" proposed by Dawkins and Hull among others, and takes the "classical approach", embraced for instance by Lewontin, as the starting point of his own proposal of an abstract representation for a Darwinian populational dynamics, that might be applied to different kinds of systems, not restricted to the biological realm (Godfrey-Smith, 2009, p. 31-6).

representation is also used to depict evolutionary transitions as well, as being trajectories in that space. Different kinds of Darwinian populations, associated with different kinds of individuals, are located in different places in the *Darwinian hyperspace* (as I will, henceforth, be calling this representation) given the values these populations score in a set of parameters that are briefly described below:

H - fidelity in inheritance

C - continuity[2]

S - relationship between fitness and intrinsic properties

V - abundance of variation

α - reproductive competition[3]

Besides those, Godfrey-Smith emphasizes the relevance of three reproduction-related parameters (see Figure 1), summing up an eight-dimensional hyperspace:

B - bottleneck

G - reproductive specialization of the parts in a collective entity[4]

I - overall integration of the collective entity

[2] The meaning of the parameter *C* can be grasped by using the idea of a *fitness landscape*. If it is *rugged*, small variations in the system's properties lead to big variations in fitness.This situation corresponds to a *low* value of the parameter C; in a landscape like this, the population can be easily trapped in a local fitness *peak* and not be able to cross a valley and to evolve towards a higher fitness *peak* on the landscape. The way the population might possibly evolve is, in this case, not *C*ontinuous, being as a result more susceptible to drift.

[3] The parameter α measures the degree in which the reproductive success of one individual in a population affects the reproductive success of another one in the same population.

[4] The parameter *G* is modeled on the *G*erm/Soma reproductive specialization in multicellular organisms.

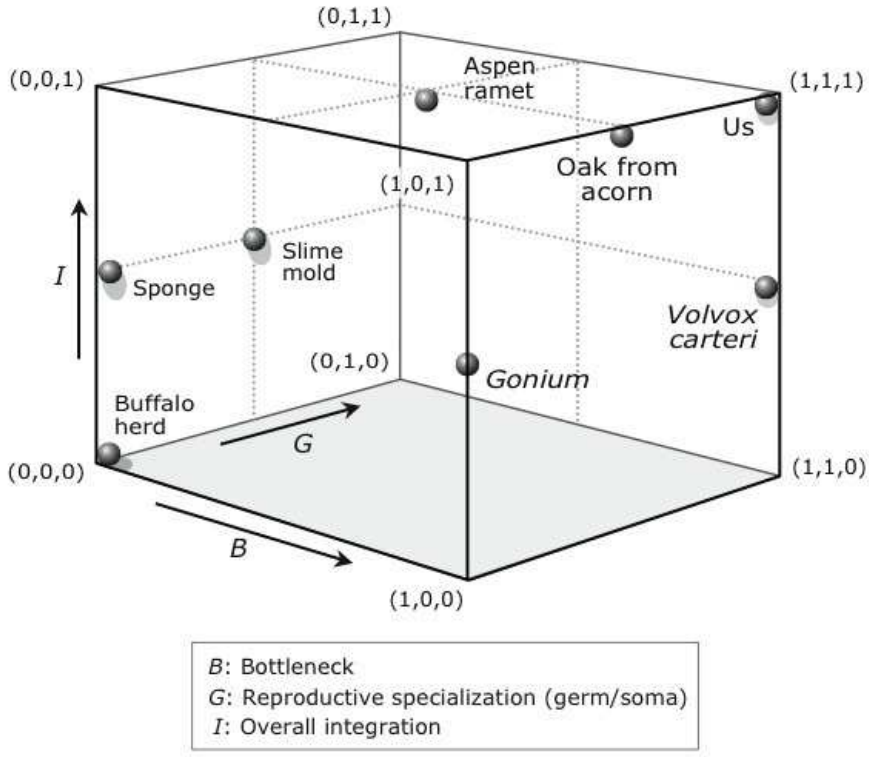


Figure 1: The Darwinian hyperspace with just three dimensions representing the reproduction-related parameters B, G and I. Several organisms are located in this space given their coordinates along these dimensions (From Godfrey-Smith, 2009, p. 95).

In the framework proposed by Godfrey-Smith, Darwinian populations have ontological priority, so to speak, vis-à-vis Darwinian individuals: "...the population-level concept comes first" (2009, 6). Therefore, any attempt to apply those parameters to track possible transitions in individuality associated with cultural change has first to address the question about what kinds of Darwinian populations might exist in this domain. This is the main topic of the next section. Afterwards, I will evaluate if it is fruitful to apply Godfrey-Smith's representation for telling paradigmatic from marginal Darwinian

populations in cultural evolution.[5]

(II) DARWINIAN POPULATIONS IN THE CULTURAL DOMAIN

Godfrey-Smith argues that there are "several ways" in which *Darwinian* populations can be represented in the cultural realm (2009, 151). He distinguishes two "options" I will be naming in this paper 'BP' and 'CP'. They are first presented in an 'individualistic' way (BP_i and CP_i). Godfrey-Smith suggests that there are also group-level descriptions (BP_g and CP_g) of Darwinian populations in this domain (see Table 1): "... we have two cross-cutting distinctions, one concerning the type of thing that makes up the population, and hence the associated notion of reproduction, and the other concerning the level at which the population exists" (Godfrey-Smith, 2009, 151).

Level / Type of thing	Individualistic	Group-level
BP (biological)	BP_i - agents having cultural phenotypes	BP_g - groups having cultural phenotypes
CP (cultural)	CP_i - cultural variants (memes)	CP_g - cultural variants' bundles (memeplexes)

Table1- Darwinian populations in the cultural realm

Individualistic descriptions

BP_i) In this option, the population is made up of "ordinary biological individuals" with different *cultural phenotypes*. Reproduction in this case is ordinary biological reproduction:

[5] This paper is part of a larger project in which I am attempting to figure out how fruitful might be to apply the whole set of parameters of Godfrey-Smith's Darwinian hyperspace to track a possible transition in individuality that could have happened in the Hominin lineage, associated with cultural change.

"When people reproduce, their offspring often resemble the parents with respect to these features, as a consequence of teaching and imitation (...) It is not a *new* application of the theory, in fact, but an ordinary one" (Godfrey-Smith, 2009, 150).

As far as inheritance is concerned, in the BP_i case we have just vertical transmission of cultural variants (or memes, if you like), through teaching and imitation.[6]

CP_i) In the second individualistic option, cultural variants themselves make up a (Darwinian) population. In the previous BP_i option, the population is made up by *the bearers* of cultural variants. Here, cultural variants themselves make up the focal population and there is replication of cultural variants. I will come back later to the modality of reproduction associated with a CP_i-like population.

Group-level descriptions

The "two options" previously described are individualistic in character but Godfrey-Smith makes explicit that there are group-level populations, as well, of biological and cultural "types of things":

"It could be argued that human groups have cultural phenotypes that are transmitted to offspring groups (...), or that group-level cultural variants themselves (such as forms of political organization) may make up a pool of reproducing entities" (Godfrey-Smith, 2009, 151).

We end up with four kinds of populations in the cultural realm: at an 'individualistic' level, the populations are either composed of biological individuals (agents, for short) with cultural phenotypes (BP_i) or made up by the cultural variants themselves (CP_i). At the group-level, either groups of agents with different cultural phenotypes (cultural groups, for short) make up the population (BP_g), or bundles of cultural variants (something akin to what memeticists call "memeplexes") themselves constitute the population (CP_g).

One might ask whether the kinds of Darwinian populations in each of the four cases (BP_i,

[6] I prefer to use the expression 'cultural variants' that is more neutral, not committing myself to the properties usually attributed to memes.

BP_g, CP_i, CP_g), admitting individualistic and group-level descriptions, are paradigmatic or marginal. To address this question we should locate each case in the proposed Darwinian hyperspace.

In the following, I will focus on the BP_g case. The chief question I want to address is whether this group-level population is paradigmatically Darwinian or just marginal.

After presenting the BP_g option, Godfrey-Smith mentions Henrich and Boyd's 1998 paper on the role played by a conformist bias in human evolution. I highlight this reference here because this transmission bias will be discussed at length in this paper.

The BP_g kind of Darwinian population is central to Richerson and Boyd's theory of human evolution, a particular brand of gene-culture coevolution theories. My bet is that their "dual inheritance" theory helps to shed light on some of the topics Godfrey-Smith addresses in his book, related to cultural evolution. And the other way around: Godfrey-Smith's way to represent transitions in individuality as trajectories in an abstract Darwinian hyperspace helps to develop further some aspects of Richerson and Boyd's theory.[7]

(III) HOW IS BP_g LOCATED IN THE DARWINIAN HYPERSPACE?

Taking for granted the conceptual framework presented above, I want to put forth once more the chief questions I will be addressing in this paper: Might human groups with different cultural phenotypes be Darwinian individuals? Do we have in BP_g a paradigmatic or a marginal Darwinian population?

To tackle these questions, we must apply Godfrey-Smith's procedure, that is, we must locate the BP_g population in the Darwinian hyperspace, by roughly indicating its *coordinates* along the eight dimensions presented above. This is a much bigger project than I will be able to accomplish in this paper. I will focus here on just a few of those parameters

[7] Boyd and Richerson share with Godfrey-Smith, furthermore, some more general points of view that invite the kind of approximation between their work I am exploring in this paper. First of all, they agree in pointing to population thinking as the most central aspect of Darwinism. They are also suspicious about the replicator approach (especially in the cultural domain) and argue that replicators are not necessary for evolution by natural selection to take place. They all embrace also a multilevel approach to natural selection.

(and point to some relationships between them) and look at how cultural groups fare in these dimensions of the Darwinian hyperspace.

De-Darwinization in BP_g

The parameter *V* measures the *abundance of variation* in a population. How much variation, and of what kind, is required to fuel an evolutionary dynamics at the level of *groups* of an BP_g type? Since we are dealing here with collective entities, we have to look also at how the population *inside* each group fares regarding the *V* parameter. In the human case, at least, the relevant lower level population is made up of agents exchanging cultural information in a social network. How abundant is the variation at this lower level, compared to the variation we find in the population of cultural groups?

If we take as a model the already mentioned case of multicellularity, there is a suppression of variation at the lower level population of cells that make up the organism: they are very similar in their intrinsic, genotipic properties.[8]

In the case of collectives, Godfrey-Smith describes an evolutionary transition as a combination of processes taking place simultaneously in nested populations, at several levels, that constitute the new individual. The evolutionary trajectory that represents the emergence of a new paradigmatic Darwinian population at the level of *collectives* in the hyperspace, and the simultaneous trajectory taken by the population of *members* of these collectives run in opposite directions.

Using Godfrey-Smith's expression, those members are "de-Darwinized" in different aspects, including *V*. In other words, in a major transition, the lower level population

[8] The lower level populations in multicellular organisms, taken as a model for a collective entity, have also other features I will not be fully addressing in this paper: a) there is a *division of labor* between somatic and reproductive parts (cells, in this case); b) the latter are *sequestered* very early in the development of the organism and, therefore, are shielded from the evolutionary activity that happens in the population of somatic cells during the life of the organism; c) there is often a *bottleneck* in the modality of reproduction they instantiate; in the clear-cut cases, the development starts with a single cell, a condition that scores the highest value in the parameter ($B = 1$), and this is the reason why the population is quite uniform in their intrinsic properties (genotypic, in the multicellularity case).

usually changes its status from paradigmatic to marginal when the transition concludes.

Reasoning the same way in the case of a population of *agents* making up a cultural group, we should expect that this population is, to some extent, de-Darwinized in the transition towards a paradigmatic Darwinian population of cultural *groups*.

If we focus on the parameter *V*, when an evolutionary transition is achieved the population of group-member agents displays less variation (in the agent's intrinsic properties), compared to the population of cultural groups.[9]

Why should we expect de-Darwinization of the lower level population when it comes to group-level phenomena? There is always the risk of subversion, by free-riders, of the cooperation and division of labor that maintains the integrity of the group (Godfrey-Smith, 2009, 101; 123). Therefore, mechanisms for leveling the fitness of altruists, on the one hand, and the fitness of selfish agents, on the other hand, have to be put in place for cooperation to be preserved.

Furthermore, variation at the group-level should be enhanced and kept (despite migration

[9] One might ask what would be *intrinsic* properties in BP_g-like populations (at the low and high levels). This is relevant for the definition of the parameter *S*, as described by Godfrey-Smith (see above). This issue is not my focus in this paper and I will just offer some crude intuitions here. At the level of groups with different cultural phenotypes, we would expect, in an evolutionary transition, that these groups score higher values in the parameter *S* as well, that is, that their fitness becomes (more) related to their intrinsic properties (in other words, that their fitness *S*upervenes on the latter properties when the transition concludes). Maybe, it is better to say that group-level intrinsic properties *emerge* in an evolutionary transition (the same for fitness as a property at this level). It is plausible, therefore, to consider those cultural variants that distinguish a group phenotype from that of another group as being *intrinsic* properties of that group. If conformism and other biases are in place - as well as moral aggression and other mechanisms for suppressing cultural variation -, we have, as a consequence, a fairly uniform population at the level of the group-members' population. In a transition, we expect that the fitness of a *group-member* will be increasingly dependent on the fitness of the cultural group, what can possibly be interpreted as a suppression of *S* at the level of the group-member's population (since location in a particular group can be interpreted as an extrinsic property of a group-member). Much more has to be done to establish fruitful relations between *S*, *V*, *H* and the reproduction-related parameters for each level in an BP_g-like collective entity.

etc.) for group selection to have strength, at the same time that (behavioral) variation inside the groups has to be suppressed.

What is at stake is the intensity of selection at the cultural group-level, which arguably has been non-negligible in human evolution at least. In Richerson and Boyd's dual inheritance theory for human evolution, psychological biases like conformism play a central role in supressing variation inside each cultural group, at the same time that these biases increase variation between these groups and maintain this variation along the time. These processes, going on simultaneously at both levels, would achieve a transition towards a Darwinian population of groups with different cultural phenotypes (BP_g).[10]

We are touching here upon the problem of the evolution of cooperation, also discussed by Godfrey-Smith (2009, p. 115;163-4). What would be the analogues, in the cultural domain, of the ways of avoiding subversion we find in the biological domain?

We know that just kin selection and reciprocal altruism are not enough to support cooperation in groups whose members are not genetically-related and/or in large groups.[11] Richerson and Boyd point, therefore, to other mechanisms of "variation supression" (to use Godfrey-Smith's expression): moralistic aggression and symbolic markers. Through these mechanisms, cultural groups score higher values in the parameter *V* and selection at the group-level becomes stronger. By the same token, cultural groups achieve a tight integration, that is, they score higher values in Godfrey-Smith's parameter *I*.[12]

How could those mechanisms for promoting cooperation in human groups have evolved?

---

[10] Besides the conformist bias, Boyd and Richerson argue for the relevance of other transmission biases in the transmission of cultural variants: the model bias and the content bias. We discuss at length the role these biases play in their theory in Abrantes & Almeida, 2011.

[11] Richerson & Boyd, 2005; Abrantes & Almeida, 2011. Cf. Godfrey-Smith, 115.

[12] The former discussion suggests that through moral aggression each group "takes control over the lives and activities of [cultural agents, in this case], especially with respect to their reproduction" (2009, 124). This is one of the ways, pointed out by Godfrey-Smith, in which lower level populations in collectives are de-Darwinized (in their reproductive output also). I am not sure whether he would accept this interpretation of the quoted passage in the context of BP_g-like populations.

Even though this question will not be thoroughly addressed in this paper, I will say a few other things on the role of transmission biases in the next section.[13]

(IV) RULES FOR UPDATING BEHAVIOR AND DARWINIAN POPULATIONS

In the chapter on "Cultural evolution" of his 2009 book Godfrey-Smith engages himself in modeling the dynamics of a population of behaviors when a particular rule, among several possibilities, is followed by the agents for updating their behavior (2009,159-60). He investigates, especially, the evolutionary implications of the following rules that might be used in this context: 'imitate your best neighbor' (IBN), 'copy the common' and 'best response'.

An agent that follows the IBN rule looks around his or her neighbors (in a local interaction) and compare their behaviors for their payoffs; the agent then chooses to imitate the behavior that gets effectively the highest payoff. A best response rule is "smarter" than IBN since the agent not only looks around for her neighbors actual behaviors but is able to find out what would have been the most appropriate behavior given their circumstances. The agent embraces the behavior that, in Godfrey-Smith words, "would have been the most appropriate overall response to the behaviors produced by the individual's neighbors on the previous time-step" (2009,157).

The 'copy the common' rule is a kind of conformist rule: the agent imitates the behavior that is more common among those to which it is exposed.

Godfrey-Smith argues that IBN can support a Darwinian dynamics in the population of behaviors, but not the 'copy the common' rule. His argument is based on two assumptions:

---

[13] The emphasis Godfrey-Smith puts on integration (the parameter *I*) in his account of the requirements for a paradigmatic darwinian population, can contribute to develop further dual inheritance theories. In my view, Hodgson and Knudsen (2010, p. 163-4) rigthly point out that a concern with social structure is lacking in Richerson and Boyd`s theory, for instance, and that we need more than psychological biases to deal adequately with the problem of the evolution of cooperation in human social groups. For an in depth discussion of the issue of cooperation, in the context of dual inheritance theory, see Abrantes & Almeida, 2011.

1) IBN is success-driven but not conformism. After all, in the first case the agent imitates the behavior that gets the highest payoff among those to which it is exposed. An agent that conforms is not, for whatever reason, in a position to evaluate the payoffs of the behaviors to which it is exposed, since the most common behavior is not necessarily the fitter one given the circumstances.[14] If we accept this assumption, IBN would be a "smarter" rule than the copy the common rule.

2) If the agents in the population follow the 'copy the common' rule, then we can't expect a Darwinian dynamics in the population of behaviors, since the behaviors that are imitated by the agents do not have single 'parent' behaviors. This rule does not give rise, therefore, to a lineage of behaviors: "... any given behavior will not have a single 'parent' behavior on the previous time-step" (Godfrey-Smith, ibid.,157).

He argues that if the IBN rule is followed instead by the agents,

"A particular instance of a behavior might, through successive events of imitation, be the ancestor of a branching tree of descendant behaviors, spreading through the population. Each behavioral instance is transitory, but if successful it may be causally responsible for other behaviors of the same kind. Behaviors themselves in this system are replicators" (Ibid., 157).

Godfrey-Smith concludes, assuming (1) and (2), that a conformist rule cannot give rise to a Darwinian change in the pool of behaviors themselves (Ibid., 160).

In what follows, I will object to the first assumption of the argument reconstructed above. I will not address the second assumption, since I have not much to say about reproduction and inheritance in this paper, despite their indisputable relevance in demarcating different kinds of populational dynamics.

Another perspective on behavior updating rules

In his discussion of various rules for updating behavior, Godfrey-Smith is clearly focusing

[14] I will put aside, for now, the issue of the psychological requirements for being able to do this kind of appraisal.

just on what I have called the CP_i case (see Table 1), that is, on the effects of following a particular rule in the dynamics of a population of behaviors (or, else, on the population of cultural variants that cause these behaviors). There is, however, another perspective that can be taken into account when addressing the evolutionary effects of following these rules, by changing the focus to the BP_g case instead. What is now at stake is the evolutionary dynamics of a population of *groups* with different cultural phenotypes, whenever a particular rule is followed by the *members* of those groups.

So that groups with different cultural phenotypes make up a (less marginal) Darwinian population, the agents that are members of these groups should follow a conformist rule, contrary to Godfrey-Smith's own expectations. I antecipated the argument supporting this thesis in the last section: a conformist rule leads to higher values of *V* for the population of cultural groups.

Furthermore, I suspect that the effects of the IBN rule on the dynamics of behaviors *internal* to a particular cultural group might endanger its cohesion, which is not the case if the copy the common rule is embraced by the agents.[15] In other words, an IBN rule might lead to higher values of *V* in the population of *group-member* behaviors, whereas the copy the common rule obviously favors a lower *V* for this population.

At the same time, I argued before that a conformist rule for updating behavior conveys higher values of the parameter *V* for the population of *groups* (that is, this population becomes more diversified as far as culture is concerned). As a consequence, they become more isolated from each other, since cultural variation builds up barriers for migration (language is very effective in this regard). In addition, this situation enhances the strength

[15] Another point that can be made is that "smarter" rules such as IBN and the best response rule presuppose that the agent is able to appraise which of her neighbors' behaviors has the best payoff under the prevailing environmental conditions. Very often, however, an agent is not able to do this - to appraise whether a particular behavior, to which it is exposed, is adaptive or not - and the best bet is to imitate the most common behavior in the group. An alternative would be for the agent to rely on individual learning, which can be a very risky strategy if, for whatever reason, the environment is informationally translucent for the agent. For the distinction between informationally opaque, transparent and translucent environments, see Sterelny, 2003.

of selection at the group-level, as I had the opportunity to emphasize before.

In other words, following a copy the common rule de-Darwinizes the group-member's population, as far as the abundance of behavioral variation is concerned.[16] A conformist bias - and maybe other biases too, besides enforcement mechanisms such as moralistic aggression -, might also reduce reproductive competition among the members of a particular group: this population scores a lower value in the parameter α.[17] Therefore, we have the conditions for a more cooperative interaction between the members of a particular cultural group. Competition switches from the level of group-members to the group-level population, where V is higher. By the same token, we should also expect a stronger selection at the cultural group-level whenever a conformist bias shapes social learning at the lower level of group-members.

Godfrey-Smith (2009, p. 157-8) makes it clear that models which address behavior updating rules, such as those built by Skirms, are attempts to simulate the conditions under which cooperation could have evolved. The group-level BP_g point of view I am suggesting in this section, points to a scenario in which a conformist bias is one of the chief elements that favored the evolution of cooperation in human cultural groups. Richerson and Boyd, among others, offered reconstructions along these lines, as I mentioned before.

Concerning the issue of the evolution of rules for updating behaviors, Godfrey-Smith says

[16] Possibly we might also have a de-Darwinization not only regarding V, but also in reproduction-related parameters as well, for the group-member's population. At the same time, a transition towards a more paradigmatic population at the cultural group-level is taking place, as far as the latter parameters are concerned. To argue thoroughly for this thesis is beyond the scope of this paper.

[17] One might ask about what is being reproduced here. The CP_i and CP_g cases correspond to populations of cultural variants, therefore the latter are the entities being reproduced. Given Godfrey-Smith's distinctions between different kinds of reproducers, it would seem straightforward to classify this kind of reproduction using the categories of formal and scaffolded reproducers, but he is not clear about it (2009, p. 79, 154-5; cf. Dennett, 2011). It is even more complicated to conceive the modality of reproduction involved in the BP_g case. Godfrey-Smith claims that there is no clear-cut (paradigmatic) reproduction in this case, which implies that we can't attribute to cultural groups the status of full individuals. I will argue against this claim at the end of the paper.

in passing:

"So evolution can build agents who use social experience to influence their choices in a number of ways. It is a striking fact that some of these ways, including IBN, can generate a new Darwinian population in the pool of behaviors themselves. But evolution may or may not build such agents. And it may build them initially and then build something beyond them - suppose biological evolution produced a sequence of successively "smarter" rules in a population: first copy-the-common, then IBN, then a best-response rule. The pool of behaviors is initially non-Darwinian, becomes Darwinian, and then becomes non-Darwinian again" (Ibid., 160).

Godfrey-Smith does not develop this scenario further in his book, but I want to point out that it refers, again, to the CP_i case (see Table1).

My focus on the BP_g case points, instead, to a more constrained scenario, in which the evolution of a copy the common rule (arguably in the Hominin lineage) is much more probable than the evolution of other rules, given the environmental conditions that prevailed during the Pleistocene (Boyd & Richerson, 2005). I would guess also that an IBN rule has a higher cost for the agent in those environmental conditions.[18]

From the point of view I am taking here, a conformist rule might be success-driven, after all, and it can be shown that it is able to support, actually, a Darwinian dynamics at the biological *group*-level population (BP_g).

There is a large amount of litterature on the role conformism might have played in Hominin

[18] Besides the point I made before concerning the effects on the parameter *V* of following the IBN rule, my intuition is that, compared to the conformist rule, the costs of following the IBN rule are higher: we have to consider the cost of the psychological machinery required for the evaluation of the payoffs and, in addition, to take into account the (cost of) risk of imitating a behavior that is not the most adaptive, given the environment in which the population has been living (refer also to the point I made in footnote 15 concerning informationally translucent environments). This is a situation in which intuition can mislead and mathematical modeling is indispensable to compare the various scenarios.

groups and on the conditions under which it might have evolved.[19] According to several models built by Richerson and Boyd, among others, the evolution of imitation as a social learning modality is closely related to the evolution of a conformist rule for updating behaviors (the equivalent to what Godfrey-Smith calls a 'copy the common' rule). Social learning by imitation enhances the fitness of the agent when certain environmental conditions prevail: those conditions in which the environment is neither too unstable – which would favor, instead, individual learning – nor very stable - which would favor an innate behavior. These models give plausibility to a scenario in which a conformist transmission bias and high-fidelity imitation evolved in the very same environmental conditions. Therefore, a conformist bias has been probably selected for at the group-level, and one of its effects was a de-Darwinization of the lower level population, as I argued above.[20]

CONCLUSION

The arguments presented in the previous sections - inspired by some of the theses defended by dual inheritance theorists -, suggest that a population of groups with different cultural phenotypes might be more paradigmatically Darwinian than Godfrey-Smith is willing to acknowledge in his 2009 book. It is true, however, that the points I make in this paper are restricted to just a few dimensions of the Darwinian hyperspace. The BP_g-like population might (still) be a marginal one, as far as other dimensions of this hyperspace are taken into account, especially those quantified by the reproduction-related parameters. Godfrey-Smith is explicit about what is at stake here:

"Darwinian language is often applied to social groups and communities in such a way that the focus is on persistence of a group as contrasted with extinction, or growth as opposed to shrinkage (...) In this book I treat Darwinian processes involving growth and persistence

[19] Henrich & Boyd, 1998; Boyd & Richerson, 2005; Abrantes & Almeida, 2011. Hodgson & Knudsen argue for a replicator approach on tackling this issue (2010, esp. 140, 159-165). I emphasized at the beginning of the paper the reasons why Godfrey-Smith rejects this approach (see also 2009, p. 110-11).

[20] Another possible scenario would be one in which a conformist bias coevolved with a capacity for high-fidelity imitation. We discuss some of those models in Abrantes & Almeida, 2011; Abrantes, 2011.

without reproduction as marginal cases (...) So "cultural group selection" of a significant kind requires differential reproduction, not just differential persistence, even though the border between these is vague" (Ibid., 151-2; cf. 118-9).

Taking this stance, Godfrey-Smith is skeptical about the possibility of talking about reproduction in the case of cultural groups. My intuition, instead, is that it might be fruitful to come up with modalities of reproduction suitable to cultural groups, such as persistence. This strategy is compatible with the "permissive attitude" (2009, 91) he embraces along the book in other hard cases and concerning other parameters of the Darwinian hyperspace.[21]

Further work has to be done to argue more forcefully in favor of the thesis that the emergence of cultural groups in the Hominin lineage might have been a transition in individuality. This is an speculative scenario, albeit plausible, suggested by Godfrey-Smith's novel approach to the issue of transitions. It is an empirical matter how far we have been going along any of those possible evolutionary paths.[22]

BIBLIOGRAPHY

Abrantes, P. Methodological issues in the dual inheritance account of human evolution. In: *Darwin´s Evolving Legacy*. Martínez Contreras J. & Ponce de León A. (eds.). México: Siglo XXI - Universidad Veracruzana, 2011, p. 127-143.

Abrantes, P. ; Almeida, F. Evolução Humana: a teoria da dupla herança. In: Abrantes, P. (org.), *Filosofia da Biologia*. Rio Grande do Sul: ARTMED, 2011, p. 261-295.

Boyd, R.; Richerson, P. *The origin and evolution of cultures*. Oxford: Oxford University Press, 2005.

[21] Refer also to the above footnotes 12 and 17. For an argument along a similar line, see Dennett, 2011.

[22] I am grateful to Peter Godfrey-Smith for several conversations we had in Harvard University in 2009, which helped me to clarify and further work out some of the topics I address in this paper. Any mistakes in it are my own responsibility, of course. Versions of this paper have been presented at the 2011 ISHPSSB Meeting (Salt Lake City) and at the VII International Principia Symposium (Florianópolis, 2011). I am grateful to the Brazilian Research Agency (CNPq) for the scholarship that made possible my stay in Cambridge and my attending those Conferences.

Buss, L. W. *The evolution of individuality*. Princeton (NJ): Princeton University Press, 1987.

Dennett, D. Homunculi rule: reflections on `Darwinian populations and natural selection` by Peter Godfrey Smith. *Biology & Philosophy*, v. 26, p. 475–488, 2011 .

Godfrey-Smith, P. *Darwinian populations and natural selection*. Oxford: Oxford University Press, 2009.

Henrich, J.; Boyd, R. The Evolution of Conformist Transmission and the Emergence of Between-Group Differences . *Evolution and Human Behavior*, v. 19, p. 215–241, 1998.

Hodgson, G. M. ; Knudsen, T. *Darwin's conjecture: the search for general principles of social & economic evolution*. Chicago: The University of Chicago Press, 2010.

Maynard-Smith, J. ; Szathmáry, E. *The major transitions in evolution*. Oxford: Oxford University Press, 1997.

Michod, R. E. *Darwinian dynamics: evolutionary transitions in fitness and individuality*. Princeton (NJ): Princeton University Press, 1999.

Richerson, P. ; Boyd, R. *Not by genes alone: how culture transformed human evolution.* Chicago: The University of Chicago Press, 2005.

Sterelny, K. *Thought in a hostile world*. Malden (MA): Blackwell, 2003.

The Moral Lives of Animals
Mugla, Turkey
June 2012

1 Introduction

Do animals lead moral lives? What exactly might be meant by claiming that they do and how might we be able to establish that fact? This is the focus of the following, programmatic, paper. My aim is to establish a framework for answering these questions and suggest a direction for further investigation.

Much of the literature on animals and morals focuses on the moral *status* of animals. Do they need to be considered in *our* moral calculations and if so how? To what extent does the moral status of animals suggest or dictate human attitudes towards them and human practices with respect to them? To borrow a phrase from Peter Singer the question is should the circle that encompasses the moral community of human beings be expanded to include some if not all animals? If so, what criteria are relevant for determining who is or is not to be included in this expanded circle? Utilitarians opt for the capacity for feel pain, Kantians and neo-Kantians opt for evidence of some degree of rationality or reflective capacity and virtue theorists, I suppose, would opt for some evidence of the manifestation of virtue. The point is that each of these approaches reflects what might be called an 'anthropocentric perspective' insofar as the key underlying question seems to be what the implications of including or excluding animals in the 'moral circle' are *for us*? They are anthropocentric in another sense as well in that who counts as morally relevant is determined by criteria that are set by some understanding of human conceptions of morality.

My approach is somewhat different. The question I am interested in exploring is this: To what extent can we get a handle on the moral lives of animals from the perspective of the animals themselves? Does it make any sense and, if so, what sense, to talk of animals as leading moral lives independently of questions about how and whether to factor them into our moral deliberations? In terms of Singer's 'expanding circle' metaphor we may put the question in the following way: Is there one moral circle that encompasses all those who warrant moral consideration or are there perhaps a number of (possibly overlapping) circles centered around different focal points? Does it, for example, make sense to talk of a moral community of wolves or elephants where the norms of these communities and the criteria for membership are determined by and reflections of the social dynamics of the respective groups? In contrast to the traditional anthropocentric perspective this approach might be labeled 'speciocentric.'

The plan of the paper is as follows. Section 2 is a brief summary of the main empirical and theoretical considerations that suggest that at least some non-human animals lead moral lives that can be appropriately characterized from a speciocentric point of view. Section 3 raises two questions that need to be addressed if the project of attributing moral lives to animals is to get off the ground. Section 4 explores the sense in which animals might be construed as moral agents. This discussion draws on some recent work by Geoffrey Sayre-McCord on the nature of normativity.[i] Section 5 is a discussion of a moral version of what is known as the 'logical problem' in the theory of mind literature. This material draws on some recent work by Robert Lurz.[ii]

2. The empirical and theoretical background

Here I briefly summarizie material that is dealt with more fully in 'The moral life of animals.'[iii] The general empirical and theoretical support for attributing moral sensibilities to animals derives from three sources: evolutionary theory, neuroscience and cognitive ethology.

The argument from evolution has its roots in the work of Charles Darwin and George Romanes.[iv] The basic idea is that human beings, other mammals and even more distant lineages have a shared evolutionary history. This history records the development of the shared underlying biological mechanisms that give rise to psychological and affective states. Different lineages may manifest those characteristics in different ways but the implication is that the differences between lineages are differences in degree and not differences in kind. The attribution of mental and affective states to animals was blacked by the rise and dominance of behaviorism in the first half of the 20th century. However, developments in neuroscience in the past 40 years have led some to challenge the behaviorist paradigm that rejects all attributions of mental or affective states to animals as anthropomorphism gone wild.

The evidence from neuroscience is extensive although the implications for attributing mental and affective states to non-human animals are still somewhat controversial. Two of the major figures advancing the view that the neuroscience strongly supports the view that animals do have minds and experience affects are Paul Maclean and Jaak Panksepp. The basic idea of Maclean's 'triune brain hypothesis' is that the evolved mammalian brain can be conveniently represented as the product of 3 developmental stages: A primitive reptilian brain located in the basal ganglia, an old mammalian brain located in the limbic system, and a new mammalian brain located in the neocortex.[v] The triune brain thesis argues for deep homologies between the brains of animals and the brains of human beings. Neurological evidence points to deep structural similarities between the ancient brain systems that we share with other animals. In particular, the ancient structures are the neural source of basic qualitative *feels* or *affects*. Jaak Panksepp has identified seven primary limbic emotional action systems which, he argues, are the basis of animal responsiveness and lie at the foundation of both emotional and cognitive states. In addition to this shared affective neurostructure, he has recently argued that mammals share brain structures that constitute what he calls "proto-selves" and "core selves." Further study, he suggests, may reveal the basis for attributing a sense of self to a wide range of animals. It stands to reason, he argues, that animals with brain structures similar to those in humans not only react in ways that make them appear to have qualitative experiences similar to those of humans when the homologous brain structures are stimulated, but also that they do in fact have those experiences.[vi]

The unregenerate behaviorists among you may object that the attribution of affects to non-human animals is unjustified anthropomorphism. Frans de Waal, among others, however, argues that it is not. de Waal argues that there is a double standard at work.[vii] On the one hand, researchers take cognitive differences to justify the non-attribution of emotional and mental capacities to animals while, on the other hand, they ignore evolutionary evidence that suggests that animals and human beings have shared inherited brain structures associated with emotional and mental capacities. de Waal labels this blind spot "Anthropodenial," which he characterizes as the *a priori* rejection of the importance of the fact that although non-human animals are not human, humans are animals.

The third line of relevant scientific findings comes from investigations by cognitive ethologists. In their book *Wild Justice*, Marc Bekoff and Jessica Pierce argue from the perspective of cognitive ethology that animals exhibit behaviors that are best interpreted as manifestations of empathy, cooperation, and a sense of fairness. In essence, "animals have morality."[viii] Bekoff and Pierce understand morality to be "a suite of interrelated other-regarding behaviors that cultivate and regulate complex interactions within social groups."[ix] However, these behaviors do not constitute morality in themselves; a certain level of cognitive and emotional sophistication is necessary. Bekoff and Pierce's approach is data-driven, and they emphasize the need and importance of expanding research beyond non-human primates to other social mammals: Hunting predators such as wolves, coyotes, and lions, as well as elephants, mice, rats, meerkats, and whales, among others. In addition, they emphasize the importance of studying animals in their natural habitats and not merely in the confines of laboratories where they are often asked to perform in accordance with the interests of animal behaviorists, which may or may not reflect the interests of the animals themselves.

Where is the line to be drawn between animals that evince morality in this limited sense and those that do not? Bekoff and Pierce suggest that the line is shifting as more empirical evidence becomes available and as our philosophical understanding of what it means to be moral is modulated by reflection on the scientific data. Although their focus is on social mammals, there is a widening body of evidence that suggests that some birds have the wherewithal to constitute a moral community, in the sense of relevant emotions, co-operation, and the like.

Although they argue that the data strongly support the attribution of morality to animals, Bekoff and Pierce also argue that what constitutes morality has to be understood as species specific. Thus, what counts as morality for human beings may not apply well to wolves, for instance. Nevertheless, they argue, the fact that human standards of morality are not appropriate for wolves does not mean that wolves do not possess some sense of moral relationships that is exhibited in their own manifestations of empathy, cooperation and a sense of fairness. The net effect is that there is not one sense of moral community and that we humans, as prototypical moral agents, may expand our understanding of morality to include some organisms and exclude others. The proper way to understand animal morality, they suggest, is to see that there are a number of distinct species-specific moral communities. Within these diverse communities, what counts as moral needs to be attuned to the characteristic features of the species themselves. Indeed, even within species, different communities may develop different social practices, so that what is acceptable in one wolf pack, for example, may not be acceptable in another.

These considerations, taken together, are compelling support for the claim that at least some animals, especially the social animals, have moral lives. The evidence is compelling but not conclusive. Putting aside behaviorist qualms there are still significant hurdles to be overcome before we can be confident in concluding that animals are moral creatures in their own right. To these qualms we now turn.

3 Two questions

There is an extensive literature on the dual questions of whether animals have minds and whether, if they do, they have a 'theory of mind.' There are two fundamental issues: (1) Do animals have minds?, and (2) Given that they do, are they capable of attributing mental states to others and acting on those attributions? Parallel questions can be raised with respect to the moral

lives of animals. (1m) Do animals have moral lives, that is, are they motivated by 'moral' considerations, properly understood? (2m) Given that they are, can they attribute moral motivations to others and act accordingly?

Robert Lurz, in a recent book, has identified two fundamental issues that need to be addressed in order to be in a position to answer questions about the mindreading capabilities of animals. One is theoretical and one is experimental. Parallel issues have to be addressed in order to be in a position to answer questions about the moral lives of animals.

For our problem, the theoretical issue is this: What does it mean to attribute moral lives to animals? In particular, what does it mean to attribute moral motivations to animals? The empirical issue is this: How best can we test for the existence moral sensibilities and moral motivations in animals?

I do not have a good answer to either of these two questions but I think we can make some headway in identifying the key questions that need to be answered and in identifying what is the proper perspective for answering them.

4. Levels of agency

To the extent that we attribute psychological and moral states to animals they are, in some sense, persons and not merely biological organisms. What, then, does it mean to attribute personhood to animals? We can adopt either an anthropocentric or a speciocentric perspective on this question. From an anthropocentric point of view, something is a person if it has a sufficient number of properties that make it an entity *like us*. I am not sure what constellation of properties this includes but the fact that in some legal sense corporations can be persons shows that the applicability of the concept is not limited to living beings. The central ideas that legitimate the extension of the idea of personhood to corporations, for instance, are notions of agency and responsibility. Corporations can act as (legal) agents and can be held (legally) responsible for their actions. However, what degree of agency and responsibility they possess is conferred upon them by human beings and their social practices. What about the moral agency of animals? Is that to be construed as merely derivative as well? A speciocentric perspective would reject this way of understanding what it means for an animal to be a moral agent. To the extent that animals lead moral lives (as opposed to being merely factors in our moral calculations) we must be able to construe them as moral agents in their own right. Can this be done?

In a recent paper responding to claims in the literature that attribute moral agency to animals, Geoffrey Sayre-McCord asks 'Just what is it to be a moral agent?'.[x] In effect, what is the nature of normativity? In his analysis, Sayre-McCord identifies several levels of 'agency' where an agent is understood to be something capable of representing its environment and acting on the basis of those representations (p. 5). These are, in order of increasing sophistication, (1) 'stimulus-response agents who 'represent the world as being a certain way and then respond directly (p. 5);' (2) 'planning agents,' which are basically stimulus-response agents with the extra capacity to identify alternative courses of action and act in accordance with some plan of action. Sayre-McCord characterizes these agents as 'decision-theoretical' agents whose behaviors can be adequately modeled by decision theory (p. 5); (3) 'strategic agents' are agents who attribute

designs and plans to others and act accordingly. Their behavior can be modeled by game theory; (4) 'norm-governed agents' are 'strategic agents . . . [who] introduce rules for behavior with which they are disposed to conform and disposed to enforce in various ways (p. 7);' and finally, (5) 'rational agents,' that is, strategic norm-governed agents who are 'able to represent the different options as better or worse, as right or wrong, or as justified or not and . . . [are] able to act on the basis of such normative representations (p. 7).'

On Sayre-McCord's account, truly moral agents need to be able to have a capacity for second-order reflection on first order states. That is, truly moral agents need not only follow norms but be capable of recognizing that they are following norms and be capable of using this reflective insight to guide their actions. This is a high bar for non-human animals to pass. Whether they are capable of passing it depends on how sophisticated their mental and psychological capacities are. Many who are willing to allow that some animals have sophisticated psychological states are reluctant to attribute reflective second-order capacities to them. Sayre-McCord, for one, allows that some animals are capable of rising to the level of norm-governed agents but he resists the attributing any rational, and hence, truly moral, agency to them.

The distinction between norm-governed agents and rational agents roughly parallels Kant's distinction between acting in accordance with duty and acting *from* duty. Indeed, this is the central theme of Sayre-McCord's analysis (p. 2) . His main project is to provide a Kantian account of rational agency freed from the metaphysical baggage of Kant's own account (p. 2) For our present purposes, the question is: 'Is norm-governed agency good enough for *non-human* moral agency? It is clear that many cognitive ethologists see the structured behavior of social animals as manifesting norm governed behavior. Some, perhaps sympathetic to the idea that such behavior doesn't rise to the Kantian level of moral agency, are content to qualify such animals as 'proto-moral' beings. I don't want to haggle over labels here but merely want to suggest that the resistance to qualifying animals as 'truly' moral may reflect a subtle anthropocentric bias. If we view human morality as one manifestation of a shared evolved set of social enabling mechanisms, then the peculiar feature of rational reflectivity, as Sayre-McCord understands it, looks more like a refinement of a capacity that is shared among many lineages rather than as a defining characteristic. If so then we can tentatively accept norm-governed agency as moral agency enough and move on to the empirical question of how to establish whether any animals do live moral lives, so understood.

5. The logical problem

Turning to the question of how to empirically test whether or not animals lead moral lives, we confront what has been labeled in the mind reading literature as the 'logical problem.' The mind reading problem is this: Is there any way to empirically distinguish between (1) animals that are mind readers, that is, animals that act in light of their attribution of intentional states to others, and (2) animals that are acting on behavioral cues but who do not attribute intentional states to others? The problem arises, in part, because, in the absence of language, the attribution of mind reading to animals is determined solely by their behavioral responses to environmental situations. In a recent book, Robert Lurz argues that all previous experimental results that suggest that some animals are mind readers are compromised by a failure to rule out

the hypothesis that the observed behaviors can be explained equally well by a 'behavior-reading' hypothesis to the effect that the animals are responding to behavioral cues and are not attributing mental states to either other conspecifics or to the experimenters.[xi]

Some take this failure to be able to discriminate between cases of mind-reading and cases of behavior-reading to be an insurmountable barrier to the unequivocal attribution of mind reading to organisms that cannot communicate their thoughts and intentions through the use of language that is intelligible to us. Lurz, however, argues that it is possible to design experiments that will be able to discriminate between the two hypotheses and he proposes several, as yet untested designs, that he claims will yield different predictions depending upon whether the tested animals are mind-readers or not. I do not want to pursue this here but rather to formulate the analogous problem for determining whether or not non-human animals lives live moral lives.

The logical problem for the moral lives question boils down to this: Is it possible to experimentally distinguish between animals that are acting in accordance with moral norms and animals that are behaving *as if* they were but for whom no moral considerations, *per se*, are relevant? If we allow, for the sake of argument, that morally motivated animals are norm-governed in Sayre-McCord's sense then what we want to know is whether the animals behavior is directed by (first-order) moral motivations or whether the characterization of their behavior as norm-governed is imposed upon their behavior by the ethological investigators.

This problem dogs much, if not all, of the cognitive ethology data that suggests that many social animals exhibit behaviors that can be interpreted as a result of the animals acknowledging and enforcing social and moral norms within their respective communities. Unlike Lurz, I do not have any good sense that these alternative accounts are empirically distinguishable. If they are not, then the claim that animals lead moral lives will remain in limbo despite the suggestive evidence from evolutionary considerations and the neuroscientific data. However, I am persuaded by the work of the cognitive ethologists that any decisive conclusions one way or the other must be the result of investigations *in situ* where experiments and observations are set up to reflect the conditions and expectations of the animals under investigation and *not* the expectations of alien investigators (that is, us).

i Geoffrey Sayre McCord, "Rational Agency and the Nature of Normative Concepts," http://philosophy.unc.edu/people/faculty/geoffrey-sayre-mccord/on-line-papers/Rational%20Agency%20and%20the%20Nature%20of%20Normative%20Concepts.pdf

ii Robert W, Lurz, *Mindreading Animals: The Debate over What Animals Know about Other Minds* (Cambridge MA: MIT Press, 2011)

iii Michael Bradie, "The Moral Life of Animals," in Tom Beauchamp and R. G. Frey (eds.) *The Oxford Handbook of Animal Ethics* (Oxford: Oxford University Press, 2011)

iv Charles Darwin, *On The Origin of Species: A Facsimile of The First Edition*, edited by Ernst Mayr (Cambridge, MA: Harvard University Press, 2000); Darwin, *The Descent of Man and selection in relation to sex* (Princeton: Princeton University Press, 1981); Darwin, *The Expression of The Emotions in Man and Animals*, 3rd ed., (Oxford: Oxford University Press, 1998); George Romanes, *Animal Intelligence* (London: Kegan Paul, Trench & Co., 1882); Romanes, *Mental Evolution in Animals* (London: Kegan Paul, Trench & Co., 1885)

v Paul MacLean, *The Triune Brain in Evolution* (New York: Plenum Press, 1990).

vi Jaak Panksepp, "Affective consciousness: Core Emotional Feelings in Animals and Humans," *Consciousness and Cognition* 14 (2005): 30-80.

vii Frans de Waal, *Primates and Philosophers: How Morality Evolved* (Princeton: Princeton University Press, 2006), pp. 61f.

viii Marc Bekoff and Jessica Pierce, *Wild Justice: The Moral Lives of Animals* (Chicago: The University of Chicago Press, 2009), p. 1.

ix Bekoff and Pierce, *Wild Justice*, p. 7.

x Sayre-McCord, p. 2

xi Lurz, *Mindreading Animals*, ch. 2.

Defusing Bertrand's Paradox

Zalán Gyenis
Department of Mathematics and its Applications
Central European University
Nádor u. 9. H-1051 Budapest, Hungary
gyz@renyi.hu

Miklós Rédei
Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science
Houghton Street, London WC2A 2AE, UK
m.redei@lse.ac.uk

July 2, 2012

Abstract

The classical interpretation of probability together with the Principle of Indifference are formulated in terms of probability measure spaces in which the probability is given by the Haar measure. A notion called Labeling Irrelevance is defined in the category of Haar probability spaces, it is shown that Labeling Irrelevance is violated and Bertrand's Paradox is interpreted as the very proof of violation of Labeling Invariance. It is shown that Bangu's attempt [2] to block the emergence of Bertrand's Paradox by requiring the re-labeling of random events to preserve randomness cannot succeed non-trivially. A non-trivial strategy to preserve Labeling Irrelevance is identified and it is argued that, under the interpretation of Bertrand's Paradox suggested in the paper, the paradox does not undermine either the Principle of Indifference or the classical interpretation and is in complete harmony with how mathematical probability theory is used in the sciences to model phenomena. It also is argued however that the content of the Principle of Indifference cannot be specified in such a way that it can establish the classical interpretation of probability as descriptively accurate, predictively successful or rational.

1 The main claims

Bertrand's Paradox, published first in [3], is regarded a classical problem in connection with the classical interpretation of probability based on the **Principle of Indifference**, and it continues to attract interest [17], [22], [2], [20] in spite of alleged resolutions that have been suggested in the large and still growing literature discussing the issue ([12] and [16] are perhaps the most well-known suggestions for

resolutions; the Appendix in [16] contains a brief summary of a number of typical views of the Paradox).

It is not the aim of this paper to offer yet another "resolution" or criticize the ones available; rather, we suggest a new interpretation of Bertrand's Paradox and analyze its relation to the classical interpretation of probability. The interpretation proposed here should make clear that Bertrand's Paradox cannot be "resolved" – not because it is an unresolvable, genuine paradox but because there is nothing to be resolved: the "paradox" simply states a provable, non-trivial mathematical fact, a fact which is perfectly in line both with the correct intuition about how probability theory should be used to model phenomena and with how probability theory is in fact applied in the sciences.

The key idea of the interpretation to be developed here is that the category of probability measure spaces with an *infinite* set of random events for which a classical interpretation of probability based on the **Principle of Indifference** can be meaningfully formulated is the one in which the set X of elementary events is a compact topological group, the Boolean algebra $\mathcal{S}$ representing the set of random events is the set of Borel subsets of X and the probability measure p_H is the (normalized) Haar measure on $\mathcal{S}$. After stating the **General Classical Interpretation** in terms of the probability measure space $(X, \mathcal{S}, p_H)$ together with the **Principle of Indifference**, we will define a notion called **Labeling Irrelevance** in this category of measure spaces: **Labeling Irrelevance** expresses the intuition that the specific way the random events are named is irrelevant from the perspective of the value of their probability understood according to the classical interpretation. It will be shown that **Labeling Irrelevance** does not hold in this category of probability measure spaces and we interpret Bertrand's Paradox as stating this provable mathematical fact.

This interpretation makes it possible to formulate precisely the extra condition on re-labelings that ensures that re-labelings *do* preserve the probabilities of events; the condition is an expression of the demand that re-labelings do not affect our epistemic status about the elementary events. We also will show that the recent attempt by Bangu [2] to block the emergence of Bertrand's Paradox by requiring re-labelings to preserve randomness cannot succeed non-trivially.

The interpretation will also make it clear that Bertrand's Paradox does *not* affect the **Principle of Indifference** and does *not*, in and by itself, undermine the classical interpretation of probability – the classical interpretation, the **Principle of Indifference** and **Labeling Irrelevance** are independent ideas. This is not to say that the classical interpretation is maintainable however; the main problem with it is that it gives the impression that it is possible to infer empirically correct probabilities from an abstract principle stating some sort of epistemic neutrality. It would be a mystery if this were possible, but we will argue in the final section that this is not possible and does not in fact happen in applications of probability theory.

2 The elementary classical interpretation of probability

Betrand's Paradox appeared at a time when probability theory had already progressed from the purely combinatorial phase involving only a finite number of random events to the period when it got intertwined with calculus. This development

began in the early 18th century with the appearance of limit theorems (theorem of large numbers, Bernoulli 1713, and central limit theorem, de Moivre 1733, [8]); yet, by the late 19th century the theory had not yet reached the maturity that would have made the mathematical foundations of the theory clear and transparent. This was clearly recognized by Hilbert, who, in his famous lecture in Paris in 1900, mentioned the need of establishing probability theory axiomatically as one of the important open problems (Hilbert's 6th problem [27], [26][p. 32-36]). Hilbert's call was answered only in 1933, when Kolmogorov firmly anchored probability theory within measure theory [13]. (See [7] for the history of some of the major steps leading to the Kolmogorovian axioms.)

In the measure theoretic approach probability theory is a triplet $(X, \mathcal{S}, p)$, where X is the set of elementary random events, $\mathcal{S}$ (the set of general random events) is a Boolean σ algebra of certain subsets of X and p (the probability) is a countably additive measure from $\mathcal{S}$ into the unit interval $[0, 1]$. Typically, one also needs random variables to describe certain features of the phenomenon to be described probabilistically: A (real valued) random variable f is a measurable function f from X into the set of real numbers $\mathbb{R}$; measurability being the requirement that the inverse image $f^{-1}(d)$ of any Borel set d in $\mathbb{R}$ belongs to $\mathcal{S}$. The measurability requirement entails that the distribution of a random variable $d \mapsto p(f^{-1}(d))$ is well-defined, the distribution of f is in fact the probability measure $p \circ f^{-1}$ on $\mathcal{B}(\mathbb{R})$ defined as $(p \circ f^{-1})(d) = p(f^{-1}(d))$ for all Borel sets $d \in \mathcal{B}(\mathbb{R})$. The number $p(f^{-1}(d))$ is the probability that f takes its value in d. Note that the events also can be regarded as random variables: an element A in $\mathcal{S}$ can be identified with the characteristic (also called: indicator) function χ_A of the set A (see e.g. [19] for the mathematical notions of measure theoretic probability).

The significance of probability theory being part of measure theory is that foundational-conceptual problems of probability theory, such as Bertrand's Paradox, can best be analyzed in terms of measure theoretic concepts. With few exceptions, the papers on Bertrand's Paradox typically do not aim at providing an analysis on this level of abstraction however, and, as a result, the precise nature of the paradox remains less clear than it should be. One such exception is Shackel's paper [22], which raises the issue of "Getting the level of abstraction right" [22][p. 156] explicitly. But the level of abstraction suggested by Shackel is a bit too high. To see why, we recall first the classical interpretation of probability together with the **Principle of Indifference** in measure theoretic terms.

The elementary version of the classical interpretation of probability concerns the probability space $(X_n, \mathcal{P}(X_n), p_u)$, where X_n is a finite set containing n number of random events and the full power set $\mathcal{P}(X_n)$ of X_n represents the set of all events. The probability measure p_u is determined by the requirement that the probability $p_u(A)$ be equal to the ratio of the "number of favorable cases to the number of all cases":

$$p_u(A) = \frac{\text{number of elements in the set } \{x_i : x_i \in A\}}{n} \tag{1}$$

This is equivalent to saying that p_u is the probability measure that is uniform on the set of elementary events. While it is not always stated and emphasized explicitly, it also is part of the classical interpretation what we call here the **Interpretive Link**: that the numbers $p_u(A)$ are related to something non-mathematical. Without such an interpretive link, the classical interpretation is not an *interpretation* of probability at all: the numbers $p_u(A)$ defined by (1) are just pure, simple mathematical relations. There are two standard **Interpretive Links**: The **Frequency Link**

and the **Degree of Belief Link**. We formulate here the first only, the latter will be discussed briefly in section 7. Thus we have the following specification of the classical interpretation:

Elementary Classical Interpretation: In case of a finite number of elementary events the probabilities of events are given by the measure p_u that is uniform on the set of elementary events and (**Frequency Link**:) the numbers $p_u(A)$ will be (approximately) equal to the relative frequency of A occurring in a series of trials producing elementary random events from X_n.

Notice the future tense in the above formulation: it is this reference for future random trials that distinguishes the classical interpretation (with the Frequency Link) from the frequency interpretation, in which the ensemble of elementary random events determining A's relative frequency must be specified *before* one can talk about probabilities (cf. [25][p. 24]).

The classical interpretation so formulated is not maintainable however: simple examples (such as throwing a loaded die) show that it is only under special circumstances that $p_u(A)$ is indicative of the frequencies with which A will occur in trials. This is what the **Principle of Indifference** is supposed to express. To state this principle we reformulate first the condition (1). Let Π_n be the group of permutations of the n element set $\{1, 2, \ldots n\}$ and $\pi \in \Pi_n$ be a permutation. Then the probability measure p_u on $\mathcal{P}(X_n)$ which is uniform on X_n is determined uniquely by the condition

$$\text{for every } \pi \in \Pi_n \text{ one has: } p_u(\{x_i\}) = p_u(\{x_{\pi(i)}\}) \quad \text{for all } i \in \{1, 2, \ldots n\} \qquad (2)$$

Elementary Principle of Indifference: *If* the permutation group Π_n expresses epistemic indifference about the elementary random events in X_n, *then* the (Elementary) **Classical Interpretation** is correct.

Thus the (Elementary) **Principle of Indifference** states that the (elementary version of the) classical interpretation of probability is maintainable only if one is epistemically neutral in some sense about the elementary events. For now, we leave it open how to specify the content of the "epistemic neutrality", we will return to the issue of epistemic neutrality in section 7.

3 The general classical interpretation of probability in terms of Haar measures

Bertrand's Paradox is typically regarded as an argument against the universal applicability of the **Principle of Indifference**: Bertrand's Paradox type arguments are intended to show that applying the **Principle of Indifference** can lead to assigning different probabilities to the same event. Both the original version of the argument and the numerous simplified versions of it involve an (uncountably) infinite number of elementary random events however. But then it is not obvious at all how one can apply the **Principle of Indifference** because the formulation of it in the previous section looses its meaning if the set of elementary events is not finite: there is no permutation group in the infinite case with respect to which one could require invariance of the measure yielding the "right" probabilities; equivalently: there is no probability measure on an infinite $\mathcal{S}$ that would be uniform on the infinite set X of elementary events. What is then the **Principle of Indifference** in

connection with such infinite probability spaces? Without answering this question in suitable generality, Bertrand's Paradox cannot be properly discussed in measure theoretic concepts.

Shackel's paper [22], which aims at an analysis of Bertrand's Paradox in abstract measure theoretic terms, realizes the importance of this question but does not offer a convincing specification of the **Principle of Indifference**: Shackel just *assumes* a measure μ on $\mathcal{S}$ and stipulates that the probabilities $p(A)$ be given by μ as $p(A) = \mu(A)/\mu(X)$ ("Principle of indifference for continuum sized sets" [22][p. 159]). But there are infinitely many measures μ on $\mathcal{S}$ that could in principle be taken as ones that define a probability p. Which one should be singled out that yields a p that could in principle be interpreted as *expressing epistemic indifference* about elements in X? This crucial question remains unanswered in [22].

It is clear that without some further structure on an infinite X it is not possible to single out any probability measure on $\mathcal{S}$ and hence it is impossible to formulate an indifference principle on such a measurable space. The formulation of the **Elementary Principle of Indifference** in terms of the permutation group Π_n gives a hint about what kind of structure is needed in the more general case however: It is a natural idea to try to replace the permutation group Π_n by another group $\mathcal{G}$ to be interpreted as expressing epistemic neutrality and hope that the elements g of $\mathcal{G}$ determine a function $\alpha_g \colon X \to X$ (an action on X) in such a way that if one requires the analogue of (2) by postulating

$$\text{for all } g \in \mathcal{G} \ : \ \ p^*(A) = p^*(\alpha_g[A]) \quad \text{for all } A \in \mathcal{S} \tag{3}$$

then the above condition (3) determines a unique probability measure p^* on $\mathcal{S}$, just like in the case of a finite number of events. Problem is that for a general measurable space $(X, \mathcal{S})$ with a continuum sized X there is no guarantee *in general* that a $\mathcal{G}$ exist leading to a p^* – much less that it leads to a *unique* p^*. There is however such a guarantee under some additional assumptions: If X itself is a topological group satisfying certain conditions.

If X is a locally compact abelian topological group, or a not necessarily abelian but compact topological group, then there exists a unique (up to multiplication by a constant) measure (called: the Haar measure) p_H on (the Borel sets of) X which is invariant with respect to the group action. Furthermore, if X is compact then the measure p_H is normalized and p_H is then a probability measure. (The Appendix collects some elementary facts about the Haar measure; equation (29) in the Appendix formulates the invariance of the Haar measure precisely).

The canonical example of an unbounded Haar measure is the Lebesgue measure on the real line: the Lebesgue measure is the unique measure on the real line that is invariant with respect to the real numbers as an additive group – the group action is the shift on the real line. The same holds for the Lebesgue measure on $\mathbb{R}^n$. The normalized restrictions of the Lebesgue measure on $\mathbb{R}^n$ to bounded, compact subsets of $\mathbb{R}^n$ are thus distinguished by the feature that they originate from a shift-invariant measure; moreover, the Lebesgue measure on any interval $[a, b]$ also can be regarded as Haar measure in its own right and the same holds for sets $\times_i^n [a_i, b_i]$ in $\mathbb{R}^n$ (cf. Appendix). Both the original Bertrand's Paradox and the simplified versions of it take the normalized restriction of the Lebesgue measure to some bounded, compact sets in $\mathbb{R}^n$ $(n = 1, 2)$ as the measure that expresses the **Principle of Indifference**. This amounts to interpreting (more or less tacitly) the group that generates the Lebesgue measure as a symmetry expressing epistemic neutrality about the elementary events.

Thus in general, the group action on X determined by X itself as a group can play the role of the action of the permutation group on X_n, and the Haar measure p_H on a compact X is the analogue of the uniform distribution on X_n if a non-zero uniform distribution on the elements X does not exist, which is the case if X is an infinite set. Note that taking the Haar measure as the analogue of the uniform distribution is also justifiable using maximum entropy techniques (see [11]). In what follows, $(X, \mathcal{S}, p_H)$ stands for a probability measure space in which X is a compact topological group with continuous group action, $\mathcal{S}$ is the Borel σ algebra on X and p_H is the Haar measure on $\mathcal{S}$. In the terminology of these group and measure theoretic notions the general classical interpretation of probability and the related principle of indifference can be consistently formulated generally as follows:

General Classical Interpretation: If X is a compact topological group, then the probabilities of the events are given by the Haar measure p_H on (the Borel sets of) X and (**Frequency Link**:) the numbers $p_u(A)$ will be (approximately) equal to the relative frequency of A occurring in a series of trials producing elementary random events from X.

General Principle of Indifference: *If* X is a compact topological group and *if* the group action expresses epistemological indifference about the elementary random events in X, *then* the General Classical Interpretation is correct.

4 Labeling Irrelevance

Part of the intuition ingrained in the classical interpretation of probability is what can be called **Labeling Irrelevance**. Intuitively, the **Labeling Irrelevance** states that from the perspective of the values of the probabilities it does not matter how the events are named: re-naming them should not change their probability. To formulate this idea precisely, we need the notion of re-labeling (re-naming) first: If $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces describing the same phenomenon then the map $h\colon X \to X'$ is called a re-labeling if it is a bijection between X and X' and both h and its inverse h^{-1} are measurable, i.e. it holds that

$$h[A] \in \mathcal{S}' \qquad \text{for all} \quad A \in \mathcal{S} \tag{4}$$

$$h^{-1}[B] \in \mathcal{S} \qquad \text{for all} \quad B \in \mathcal{S}' \tag{5}$$

(Here $h[A] = \{h(x) : x \in A\}$ and $h^{-1}[A'] = \{h^{-1}(x') : x' \in A'\}$.) Note that without the measurability condition required of h it can happen that a general event $A \in \mathcal{S}$ has probability but its re-named version $h[A]$ does not – in this case h cannot be called re-naming of random events (and similarly for h').

Labeling Irrelevance is the claim that from the perspective of probabilities (understood in the spirit of the classical interpretation), naming is irrelevant; that is to say, if $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces and h is a re-labeling between X and X' then it holds that

$$p'_H(h[A]) = p_H(A) \qquad \text{for all} \quad A \in \mathcal{S} \tag{6}$$

$$p_H(h^{-1}[A']) = p'_H(A') \qquad \text{for all} \quad A' \in \mathcal{S}' \tag{7}$$

Recall (see e.g. [1][p. 3]) that two probability measure spaces $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are called isomorphic if there are sets $Y \in \mathcal{S}$ and $Y' \in \mathcal{S}'$ such that $p(Y) = 0 = p'(Y')$ and there exists a bijection $f\colon (X \setminus Y) \to (X' \setminus Y')$ such that both f and its inverse

f^{-1} are measurable and such that both f and f^{-1} preserve the measure p and p', respectively; i.e. (8)-(9) below hold:

$$p'(f[A]) = p(A) \quad \text{for all} \quad A \in \mathcal{S} \tag{8}$$
$$p(f^{-1}[A']) = p'(A') \quad \text{for all} \quad A' \in \mathcal{S}' \tag{9}$$

The function f is called then an isomorphism between the probability measure spaces. **Labeling Irrelevance** can therefore be expressed compactly by saying

Labeling Irrelevance: Any re-labeling between probability spaces $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ is an isomorphism between these probability spaces.

5 General Bertrand's Paradox

Labeling Irrelevance is obviously a very strong claim and Bertrand's paradox can be interpreted as the proof that it cannot be maintained in general (see below). But why would one think that **Labeling Irrelevance** holds in the first place? The answer is: because **Labeling Irrelevance** *does* hold for an *infinite* number of probability spaces: for probability spaces with any *finite* number elementary events probabilities of which are given by the uniform probability measure. A bijection h between two finite sets X_n and $X' = X_m$ of elementary events exists if and only if the sets X_n and X_m have the same number of elements, $n = m$, and this entails that the two uniform distributions on those equivalent sets will assign the same probability to A and $h[A]$ (and to A' and $h^{-1}[A']$) – no Bertrand's Paradox can arise in this case. Since the intuition about probability theory was shaped historically by situations involving only a finite number of random events, it is not surprising that **Labeling Irrelevance** became part of the intuition about probability. It turns out however that this intuition is a poor guide if the set of elementary events is not finite: This is precisely what Bertrand's Paradox shows, general form of which is the following statement:

General Bertrand Paradox: Let $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ be probability spaces with compact topological groups X and X' having an infinite number of elements and p_H, p'_H being the respective Haar measures on the Borel σ algebras $\mathcal{S}$ and $\mathcal{S}'$ of X and X'. Then **Labeling Irrelevance** does not hold for $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ in the sense that

- either there is no re-labeing between X and X';
- or, if there is a re-labeling between X and X', then there also exists a re-labeling that violates **Labeling Irrelevance**.

The **General Bertrand's Paradox** is a trivial consequence of the following non-trivial theorem in measure theory:

Proposition 1 ([24], [21])**.** *If X is an infinite, compact topological group with the Haar measure p_H on the Borel σ algebra $\mathcal{S}$ of X, then there exists an autohomeomorphism θ of X and an open set E in $\mathcal{S}$ such that $p_H(\theta[E]) \neq p_H(E)$.*

By definition an autohomeomorphism θ of X is a bijection from X into X such that both θ and its inverse θ^{-1} are continuous. Since continuous functions are Borel measurable, an autohomeomorphism is a re-labeling: a re-labeling of X in terms of its own elements. Assume now that $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces with infinite, compact topological groups X and X' and Haar measures p_H

and p'_H. If $h\colon X \to X'$ is a re-labeling between X and X' then either h is an isomorphism between the probability spaces (i.e. preserves the probability in the sense of (6)-(7)) or it is not. If it is not, then **Labeling Invariance** is violated by h. If h *does* preserve the probability (and is thus an isomorphism between $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$) then by Proposition 1 there exists an autohomeomorphism θ on X and there exists an open set $E \in \mathcal{S}$ such that $p_H(\theta[E]) \neq p_H(E)$. This means that for the re-labeling given by the composition $h \circ \theta$ we have

$$p'_H((h \circ \theta)[E]) = p'_H(h[\theta[E]]) = p_H(\theta[E]) \neq p_H(E) \tag{10}$$

so the re-labeling $h \circ \theta$ violates (6) and thus $h \circ \theta$ violates **Labeling Invariance**. In either case **Labeling Invariance** is violated. Furthermore, the autohomeomorphism ensured by Proposition 1 provides a re-labeling of the elementary set of events of any infinite compact group in terms of its own elementary events in such a way that the Haar measure yielding the probabilities of the events in the spirit of the classical interpretation are not preserved under the re-labeling.

The General Bertrand's Paradox is thus a general feature of infinite probability measure spaces with the Haar measure yielding the probabilities, and note that it says more than the original Bertrand's Paradox, which only claimed that there exist Haar measures and re-labelings that violate **Labeling Irrelevance**: The General Bertrand's Paradox says that *no two* Haar probability spaces can satisfy **Labeling Irrelevance**; i.e. if there is at all a re-labeling between two probability spaces $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ with infinite X and X' then there is also re-labeling between these spaces that violates **Labeling Invariance**, and for *any* space $(X, \mathcal{S}, p_H)$ with an infinite X there exists a space (namely itself) and a self-re-labeling of $(X, \mathcal{S}, p_H)$ that violates **Labeling Invariance**. Thus Bertrand's 1888 Paradox can be viewed as the specific "Lebesgue measure case" of a mathematical theorem that was proved in full generality in 1993 only.

We close this section by giving an explicit, elementary example of violation of **Labeling Invariance**; this example will be referred to in the next section. In a well-defined sense (explained in Remark 1) the example is general.

Example Let $[a, b]$ and $[c, d]$ be two closed intervals of the real numbers and

$$([a, b], \mathcal{S}_{[a,b]}, p_{[a,b]}) \text{ and } ([c, d], \mathcal{S}_{[c,d]}, p_{[c,d]})$$

be the two probability spaces with $p_{[a,b]}$ and $p_{[c,d]}$ being the normalized Lebesgue measures on the intervals $[a, b]$ and $[c, d]$, with $\mathcal{S}_{[a,b]}$ and $\mathcal{S}_{[c,d]}$ being the Borel measurable sets of the respective intervals. Elementary algebraic calculation and reasoning show that one can choose the parameters α, β and γ in the definition of the simple quadratic map h defined on the real line by

$$h(x) = \alpha x^2 + \beta x + \gamma \tag{11}$$

in such a way that h maps $[a, b]$ to $[c, d]$ bijectively and both h and its inverse are continuous hence (Borel) measurable. Thus (the restriction to $[a, b]$ of) h is a re-labeling between $([a, b], \mathcal{S}_{[a,b]}, p_{[a,b]})$ and $([c, d], \mathcal{S}_{[c,d]}, p_{[c,d]})$. Specifically, the parameters below have this feature

$$\alpha = \frac{d-c}{(b-a)^2} \tag{12}$$

$$\beta = -2a\frac{d-c}{(b-a)^2} \tag{13}$$

$$\gamma = a^2\frac{d-c}{(b-a)^2} + c \tag{14}$$

Furthermore, if ϵ is a real number such that $[a, a+\epsilon] \subseteq [a, b]$ then

$$p_{[a,b]}([a, a+\epsilon]) = \frac{\epsilon}{b-a}$$

and since h takes $[a, a+\epsilon]$ into $[c, c+\frac{d-c}{(b-a)^2}\epsilon^2]$ one has

$$p_{[c,d]}(h\big[[a, a+\epsilon]\big]) = \frac{1}{d-c}\Big(c + \frac{d-c}{(b-a)^2}\epsilon^2\Big)$$

It is clear then that for many ϵ

$$p_{[a,b]}([a, a+\epsilon]) = \frac{\epsilon}{b-a} \neq \frac{1}{d-c}\Big(c + \frac{d-c}{(b-a)^2}\epsilon^2\Big) = p_{[c,d]}(h\big[[a, a+\epsilon]\big]) \qquad (15)$$

which is a violation of **Labeling Irrelevance**.

Remark 1. Note that the above example is typical in the following sense: A probability measure space is called a *standard probability space* if X is a complete, separable metric space and $\mathcal{S}$ is the Borel σ algebra of X. Standard, non-atomic probability spaces are isomorphic to $([a, b], \mathcal{L}_{[a,b]}, p_{[a,b]})$ with some interval $[a, b]$ where $\mathcal{L}_{[a,b]}$ is the algebra of Lebesgue measurable sets in $[a, b]$ (see [1][Chapter 1, p. 3]). Hence the above example gives a large number of re-labelings that violate **Labeling Irrelevance** in the category of spaces $(X, \mathcal{S}, p_H)$ with X being a complete, separable metric space. This covers all the spaces that occur in connection with Bertrand's Paradox.

6 Attempts to save Labeling Irrelevance

One may attempt to defend **Labeling Irrelevance** by trying to block the emergence of Bertrand's Paradox. The previous section makes it clear what the possible strategies are to achieve this: One can impose some extra condition on re-labelings that entails either that re-labelings satisfying the extra conditions do not exist (Strategy A) or that the re-labelings satisfying the additional conditions force the re-labelings to be isomorphisms of the probability spaces (Strategy B). Although not formulated in this terminology, Bangu's recent attempt [2] is an example of Strategy A. We show below that Bangu's suggestion for Strategy A is ambiguous however and that resolving the ambiguity makes it either a trivial case of Strategy B or is unsuccessful. A *successful* implementation of Strategy B is to say that it is unreasonable to expect a re-labeling to preserve probabilities unless the re-labeling also preserves our epistemic status with respect to the elementary events: after all, the **Principle of Indifference** states that p_H is the correct probability *only if* the group structure of X expresses epistemic neutrality. So the following stipulation is in the spirit of the **Principle of Indifference**:

Definition: The re-labeling h between probability spaces $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ preserves the epistemic status if it is a group isomorphism between X and X'.

Since the probability measures p_H and p'_H are completely determined by the respective group actions, re-labelings that preserve the epistemic status are isomorphisms between the measure spaces, hence no Bertrand's Paradox can arise with respect to such re-labelings; furthermore, not every re-labeling is a group isomorphism – thus this strategy works in a non-trivial way.

Bangu's suggestion is that one should only expect **Labeling Irrelevance** to hold for bijections that "preserve randomness" – this is his **Assumption R** – Bertrand's paradox is only a paradox in his view if **Labeling Irrelevance** is violated by re-labelings satisfying the randomness condition, which, he claims, has *not* been shown and the burden of proof is on those who claim such re-labelings exist. It is clear from the wording of his paper that he conjectures that no such proof can be given, i.e. that no randomness preserving re-labelings exist that violate **Labeling Irrelevance** (i.e. that he is following Strategy A).

As Bangu also points out, the notion of randomness is notoriously both vague and rich: the adjective "random" can be applied to different entities (events, processes, dynamics, ensembles etc.), it can come in the form of a pre-theoretical informal intuition, in form of precise mathematical definitions, and it also can come in degrees. Thus one has to be very careful and specific when it comes to the problem of whether "randomness is preserved" under a re-labeling of the elementary events. Bangu leaves it deliberately open in what sense precisely "randomness" might not be invariant under re-labeling of the random events; hence his suggestion remains somewhat vague. No matter what kind of notion of randomness one has in mind, if it is to be relevant for probabilistic modeling of a phenomenon, then it must be expressible in terms of probabilities, since the basic principle guiding the modeling of phenomena by probability theory is the maxim:

Distribution Relevance: "A property is probability theoretical if, and only if, it is describable in terms of a distribution" [15][p. 171].

In the spirit of **Distribution Relevance** one can take the position that randomness of a phenomenon expressed by "randomness" of the random variables that describe the phenomenon are encoded in the *distribution* of the random variables. Consequently, under this interpretation of randomness, if one is given two probability models $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ of a given phenomenon and $h\colon X \to X'$ is a re-labeling between $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$, then h preserves the randomness of the two probabilistic descriptions if and only if it holds that if $f\colon X \to \mathbb{R}$ is any random variable in $(X, \mathcal{S}, p)$ with distribution $p \circ f^{-1}$ then the distribution $p' \circ {f'}^{-1}$ in $(X', \mathcal{S}', p')$ of the re-named random variable $f' = f \circ h^{-1}$ coincides with $p \circ f^{-1}$:

$$\big(p' \circ (f \circ h^{-1})^{-1}\big)(d) = \big(p \circ f^{-1}\big)(d) \qquad \text{for all } d \in \mathcal{B}(\mathbb{R}) \tag{16}$$

and conversely: for every random variable $g'\colon X' \to \mathbb{R}$ which is the re-named version of a random variable $g = g' \circ h$ in $(X, \mathcal{S}, p)$ it holds that the distribution $p' \circ {g'}^{-1}$ in $(X', \mathcal{S}', p')$ of g' and the distribution $p \circ g^{-1}$ of $g = g' \circ h$ in $(X, \mathcal{S}, p)$ coincide:

$$(p \circ (g' \circ h)^{-1})(d) = (p' \circ {g'}^{-1})(d) \qquad \text{for all } d \in \mathcal{B}(\mathbb{R}) \tag{17}$$

Since the random events themselves are random variables, the two equations (16)-(17) must hold for every characteristic function χ_A ($A \in \mathcal{S}$) in place of f and every characteristic function $\chi_{A'}$ ($A' \in \mathcal{S}'$) in place of g' as well, so this requirement of preserving randomness amounts to the demand that the following two equations hold:

$$p'(h[A]) = p(A) \qquad \text{for all} \quad A \in \mathcal{S} \tag{18}$$

$$p(h^{-1}[A']) = p'(A') \qquad \text{for all} \quad A' \in \mathcal{S}' \tag{19}$$

which is precisely **Labeling Irrelevance** (eqs. (6)-(7)). So, if "preserving randomness by re-labeling" in **Assumption R** is understood in the spirit of **Distribution**

Relevance as conditions (16)-(17) then the only randomness-preserving re-labelings are the isomorphisms and no Bertrand paradox can arise indeed – requiring preserving randomness in this sense is equivalent to the requirement that the re-labelings are isomorphism, Strategy A, so interpreted, is trivial.

One can try to argue that this is an extremely strong interpretation of "preserving randomness" and that randomness also can be interpreted differently as expressed by some other property $\Phi(p)$ of the probability measure p. For instance, one has the intuition that a probability measure sharply concentrated on a single point in X is far less "random", it represents much more certainty by having zero variance than a probability distribution that has a large variance. The usual (Shanon) entropy of a probability measure also can be taken as a measure of "randomness" of the phenomenon that the probability model describes [4][p. 61-62]. Thus one can interpret the requirement of "preserving randomness under re-labeling" in **Assumption R** in different ways depending on what property Φ one chooses:

Assumption R[Φ]: If $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces and h is a re-labeling between X and X' then we say that **Assumption R**[Φ] is satisfied if both $\Phi(p_H)$ and $\Phi(p'_H)$ hold.

It is clear then that if there is a property Φ of randomness of a probability measure and there exists probability spaces $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ with a re-labeling $h\colon X \to X'$ such that **Assumption R**[Φ] is satisfied but **Labeling Irrelevance** is violated by h then Bertrand's paradox re-emerges.

The variance and the entropy are such properties: Consider the probability spaces $([a,b], \mathcal{S}_{[a,b]}, p_{[a,b]})$ and $([c,d], \mathcal{S}_{[c,d]}, p_{[c,d]})$ described in the **Example** in section 5. The variance $\sigma(p_{[a,b]})$ of the normalized Lebesgue measure $p_{[a,b]}$ on any interval $[a,b]$ is by definition equal to

$$\sigma(p_{[a,b]}) = \int_a^b \frac{1}{b-a} x^2 dx - \Big[\int_a^b \frac{1}{b-a} x dx\Big]^2 = \frac{(b-a)^2}{12} \tag{20}$$

and the entropy $E(p_{[a,b]})$ of $p_{[a,b]}$ is by definition

$$E(p_{[a,b]}) = -\int_a^b x \log(x) dx = \log(b-a) \tag{21}$$

It follows then that if $b-a = d-c = t$ then

$$\sigma(p_{[a,b]}) = \sigma(p_{[c,d]}) \;=\; \frac{t^2}{12} \tag{22}$$

$$E(p_{[a,b]}) = E(p_{[a,b]}) \;=\; \log(t) \tag{23}$$

On the other hand, the map h defined in the **Example** remains a re-labeling even if $b-a = d-c$ and **Labeling Irrelevance** is violated by this map because for $b-a = d-c = t$ eq. (15) entails that for many ϵ we have

$$p_{[a,b]}([a, a+\epsilon]) = \frac{\epsilon}{t} \neq \frac{1}{t}\Big(c + \frac{1}{t}\epsilon^2\Big) = p_{[c,d]}(h\big[[a, a+\epsilon]\big]) \tag{24}$$

Thus Bertrand's paradox re-emerges: The probability space $([c,d], \mathcal{S}_{[c,d]}, p_{[c,d]})$ can be regarded as a re-named version of the probability space $([a,b], \mathcal{S}_{[a,b]}, p_{[a,b]})$ via the re-labeling h defined by (11) and (12)-(14), furthermore, if $b-a = d-c$ then this re-labeling satisfies **Assumption R**[Φ] with Φ being the variance or entropy, and because of (15) h violates **Labeling Irrelevance** (6)-(7).

One also can try to question **Distribution Relevance**. But if one gives up **Distribution Relevance** and interprets "randomness" in a way that makes randomness not expressible exclusively in terms of the distributions involved, then the appropriately modified **Assumption R** constrains even less the emergence of Bertrand's Paradox. Rowbottom and Schackle [20] take **Assumption R** to be (a technically undefined) "unpredictability" and argue (informally) that there are re-labelings that preserve "unpredictability" and which are not isomorphisms, contrary to what Bangu [2] seems to conjecture. As a technically more explicit example, assume that a dynamic $\{\alpha_t : t \in \mathbb{R}\}$ is given on $([a,b], \mathcal{S}_{[a,b]}, p_{[a,b]})$ and a dynamic $\{\alpha'_t : t \in \mathbb{R}\}$ is given on $([c,d], \mathcal{S}_{[c,d]}, p_{[c,d]})$, where α_t and α'_t are one parameter groups of measure preserving maps on $[a,b]$ and $[c,d]$ respectively. As randomness of the dynamical systems $([a,b], \mathcal{S}_{[a,b]}, p_{[a,b]}), \{\alpha_t\})$ and $([c,d], \mathcal{S}_{[c,d]}, p_{[c,d]}), \{\alpha'_t\})$ one can take the randomness of the respective *dynamics* such as ergodicity, or mixing, which are not expressible in terms of $p_{[a,b]}$ and $p_{[c,d]}$ only. Given the re-labeling h between $([a,b], \mathcal{S}_{[a,b]}, p_{[a,b]})$ and $([c,d], \mathcal{S}_{[c,d]}, p_{[c,d]})$ described in the **Example** in section 5 that violates **Labeling Irrelevance** one can then specify the dynamics $\{\alpha_t\}$ and $\{\alpha'_t\}$ in such a way that they are both ergodic, [4][p. 34], generating a Bertrand's Paradox, or in such a way that $\{\alpha_t\}$ is ergodic whereas $\{\alpha'_t\}$ is not, which would be a violation of preserving randomness (**Assumption R**) hence not a case of Bertrand's Paradox (according to Bangu's requirement) – anything is possible under such a dynamical interpretation of randomness.

Thus the emergence of Bertrand's paradox cannot be blocked in a non-trivial way by requiring the paradoxical examples to satisfy the randomness test and showing that they cannot pass this test: unless one requires in effect that the re-labeling be an isomorphism, Bertrand's Paradox emerges: If **Distribution Relevance** is accepted and randomness is interpreted as measured by the variance or entropy of the probability measures then elementary examples can be given that show violation of **Labeling Irrelevance**. If **Distribution Relevance** is abandoned then the randomness requirement can be satisfied even more easily.

7 Comments on the classical interpretation

While Bertrand's Paradox shows that **Labeling Irrelevance** cannot be maintained in general, this does not undermine, in and by itself, either the classical interpretation of probability or the **Principle of Indifference**: It is clear from the discussion in the previous sections that the **Principle of Indifference** and **Labeling Irrelevance** are *independent* ideas: One can in principle maintain the classical interpretation based on the **Principle of indifference** and reject **Labeling Invariance** completely or restricting it to the domain in which it holds: in the category of probability measure spaces with a finite number of random events, or to re-labelings that preserve the epistemic status.

Thus Bertrand's Paradox is defused; however this is not to be taken as defence of the classical interpretation. The classical interpretation is deeply problematic for simple, non-technical reasons that are related to the general issue of how one should view the status of probability theory.

One has to distinguish *applications* of probability theory from *interpretations* of probability as this latter term is used in philosophy of science. Probability theory is part of pure mathematics in the first place. In an application of probability theory one relates the mathematical elements in a triplet $(X, \mathcal{S}, p)$ to non-mathematical

entities. This involves two tasks:

Event Interpretation To specify what the elements in X and $\mathcal{S}$ stand for.

Truth Interpretation To clarify when the proposition "$p(A)$=r" is true/false.

In an application, probability theory thus becomes a mathematical *model* of a certain phenomenon that is external to mathematics. A probability measure space is a good model of the phenomenon if it has two features: descriptive accuracy and predictive success. Descriptive accuracy means that under the fixed specification of the Event and Truth Interpretations propositions such as $p(A) = r$ are true about events that have been observed in the past. Predictive success means that the probabilistic propositions $p(A) = r$ will be true in future observations. It is clear that both descriptive correctness and predictive success are *robustly empirical features*; hence, whether a probability space is a good model is a question that can be answered only on the basis of empirical considerations. This is of course not new, there is nothing peculiar or mysterious about probabilistic modeling, probabilistic scientific theories are just like any scientific theory from this perspective. [1]

The mathematical notion of isomorphism between probability measure spaces is in complete harmony with the application of probability theory – and so is the General Bertrand Paradox: The Event Interpretation and Truth Interpretation are conceptually different issues, the former does not determine the latter, and, accordingly, two probability spaces are defined to be isomorphic if *two* conditions are satisfied: the random events in the two spaces are connected by a re-labeling *and* the re-labeling preserves the probabilities. From the perspective of the notion of isomorphisms of probability spaces finite probability spaces with the uniform probability measure just happen to have the "contingent" feature that in this category re-labelings *are* isomorphisms; in this case the re-labelings contain enough information to make them isomorphisms.

Interpretations of probability are typical *classes* of applications of probability theory, classes consisting of applications that possess some common features, which the interpretation isolates and analyzes. The main problem with the Classical Interpretation (understood with the amendment of the **Principle of Indifference**) is that it disregards the empirical character of the applications of probability theory and gives the impression that descriptive accuracy and predictive success in applications are based on (and can be ensured by referring to) an priori-flavored principle that expresses some sort of epistemic indifference about random events. But this is not possible, which is shown by the difficulty (often pointed out in connection with the **Principle of Indifference** [9]) that it is unclear how to specify the precise content of "epistemic neutrality" in such a way that the **Principle of Indifference** does not become circular and holds nevertheless: The **Principle of Indifference** holds only if epistemic neutrality *does* entail that the probabilities of the events given by the uniform probability measure *will* be equal to the frequencies of events in actual trials producing elementary random events, and such a conclusion cannot be validly based on a priori considerations – if it could, the **Principle of Indifference** would have solved the problem of induction.

[1] Although Marinoff [16] does not emphasize the empirical aspect of probabilistic modeling, his resolution of Bertrand's Paradox is essentially in the spirit of probabilistic modeling described here: Marinoff distinguishes different types of random generators representing different types of randomness and notes that, depending on which random generator produces the random events featuring in a Bertrand Paradox type situation, one obtains different probability distributions – there is nothing paradoxical about this.

One might say that the classical interpretation and the **Principle of Indifference** should be taken not with the Frequency Link but with the Degree of Belief Link, according to which p_H should be viewed as representing degrees of belief [5], [17]. To assess the viability of such an interpretation of the classical interpretation one has to distinguish two further specifications of the notion of degree of belief: *descriptive* and *normative*.

In the descriptive interpretation the claim is that p_H does represent the degree of belief of a particular person (or a specific group of people) about random events happening if the persons are epistemologically neutral about the events. Whatever the precise content of this epistemological neutrality, this descriptive interpretation of the degrees of belief is again an *empirical claim* about the thinking and behavior of certain people, which may or may not be true; testing it (including testing if the people in question have degrees of belief indeed) is a matter for empirical psychology – but this interpretation has little to do with how probability theory is applied in the sciences.

In the normative interpretation p_H is declared to stand for the *rational* degrees of belief of an abstract person (agent) if the agent is epistemologically neutral about the elementary events. In this case one has to ask in what sense and why p_H represents *rational* degrees of belief? One answer can be that p_H is rational if $(X, \mathcal{S}, p_H)$ is a good model of a certain phenomenon in the sense described earlier in this section and a rational agent's belief better be in harmony with the probabilities provided by a good model. This interpretation of rationality of p_H is essentially the content of the Principle Principle [14] and, while it is very natural, one should realize that p_H features in it in *two* roles: (i) standing for the degree of belief *and* (ii) representing some extra-mental, non-degree-of-belief-type quantities (for instance frequencies or some other dimensionless physical quantities [23]) with which the degrees of belief are required to be equal. Thus this interpretation reduces the Degree of Belief Link to another Interpretive Link and thereby the rationality (or otherwise) of an agent's degree of belief is made again dependent on empirical matters. But then it does not matter from the perspective of rationality of the degrees of belief whether the agent is epistemically neutral about the elementary events or not, because the correctness of the probabilistic model is an empirical matter that cannot be ensured on the basis of an a priori neutrality, and probability measures different from p_H can very well be rational if they satisfy the Principle Principle and the probabilistic model is good. Another possible specification of rationality of the agent's degrees of belief can be that they are consistent, i.e. that p_H satisfies the axioms of probability. Obviously, this does not single out p_H as the only rational probability.

In sum: Bertrand's Paradox interpreted as violation of **Labeling Irrelevance** does not undermine the classical interpretation of probability understood with the **Principle of Indifference**, and violation of **Labeling Irrelevance** is in complete harmony with how mathematical probability theory is used in the sciences to model phenomena; yet, irrespective of Bertrand's Paradox, the content of the **Principle of Indifference** cannot be specified in such a way that it can establish the classical interpretation of probability as descriptively accurate, predictively successful or rational.

Appendix

This Appendix recalls some elementary facts about the Haar measure. Standard references for the Haar measure are [18] and [10][Chapter XI.], for a more recent presentation see [6].

X is called a topological group with multiplication $(x, y) \mapsto x \cdot y$ and inverse $x \mapsto x^{-1}$ if the map $(x, y) \mapsto x^{-1} \cdot y$ is continuous $(x, y \in X)$. A measure p on the Borel algebra $\mathcal{S}$ of the group X is called *left* invariant (respectively *right* invariant) with respect to the group action if eq. (25) (respectively eq. (26)) below hold

$$p(A) = p(xA) \qquad \text{for all } x \in X \quad A \in \mathcal{S} \tag{25}$$
$$p(A) = p(Ax) \qquad \text{for all } x \in X \quad A \in \mathcal{S} \tag{26}$$

where for an $x \in X$, the sets xA and Ax are defined by

$$xA = \{x \cdot y : y \in A\} \tag{27}$$
$$Ax = \{y \cdot x : y \in A\} \tag{28}$$

The measure p is called invariant if it is *both* left *and* right invariant, i.e. if

$$p(A) = p(xA) = p(Ax) \qquad \text{for all } x \in X \quad A \in \mathcal{S} \tag{29}$$

On any locally compact topological group there exists both a left p_H^L and a right p_H^R invariant Haar measure and they are unique up to multiplication by a constant. The left and right invariant Haar measures are in general different. Since both the left and Haar measure is unique up to constant multiplication, and since for any $x \in X$ the measure $p_x(A) \doteq p_H^L(Ax)$ is again a left invariant measure, there exists a real number $\Delta(x)$ such that $p_x(A) = \Delta(x) p_H^L(A)$. The map $x \mapsto \Delta(x)$ is called the modular function of the group. If $\Delta(x) = 1$ for all x, then the groups are called *unimodular*; for unimodular groups the left and right invariant Haar measures coincide and yield an invariant measure. Compact and locally compact abelian groups are unimodular. The Haar measure is bounded if and only if X is compact – the Haar measure is then a probability measure.

The canonical examples of unbounded Haar measures are the Lebesgue measure on the real line and the Lebesgue measure on $\mathbb{R}^n$. It is shown below that the normalized restrictions of the Lebesgue measure on $\mathbb{R}^n$ to subsets of the form $\times_i^n [a_i, b_i)$ in $\mathbb{R}^n$ also can be regarded as Haar measures in their own right with respect to a compact group $\mathcal{G}$. This entails that the Lebesgue measure on the *closed* set $\times_i^n [a_i, b_i]$ also can be viewed as a Haar measure with respect to $\mathcal{G}$ because the Lebesgue measure space over $\times_i^n [a_i, b_i)$ and over $\times_i^n [a_i, b_i]$ are isomorphic. (Note that $\mathcal{G}$ is *not* the shift; it cannot be since shifted subsets of $[0, 1)$ are not necessarily subsets of $[0, 1)$ and the group of "shifts modulo 1" do not form a topological group due to discontinuity of the "shift modulo 1" operation.) Since $[0, 1)$ can be mapped onto $[a, b)$ by a continuous linear bijection connecting the (normalized) Lebesgue measures on the intervals $[0, 1)$ and $[a, b)$, to see how the Lebesgue measure on $[a, b)$ is a Haar measure in its own right, it is enough to see how the (normalized) Lebesgue measure $p_{[0,1)}$ on the interval $[0, 1)$ emerges as a Haar measure. Let

$$S^1 = \left\{ z \in \mathbb{C} : \ |z| = 1 \right\}$$

be the unit circle on the complex plane. As S^1 is a compact topological subgroup of $\mathbb{C}$ with the multiplication of complex numbers as the group operation, there exists a normalized Haar measure p_H on S^1. The exponential function f defined by

$$f : [0, 1) \to S^1, \qquad f(t) = e^{2\pi i t}$$

is a continuous and continuously invertible bijection between the unit interval $[0, 1)$ and the unit circle S^1; hence both f and its inverse are measurable. We claim that

f is a measure theoretic isomorphism between the interval $[0,1)$ with the Lebesgue measure on it and S^1 with the measure p_H on it; i.e. that

$$p_H \;=\; p_{[0,1)} \circ f^{-1} \tag{30}$$

To verify (30), by the uniqueness of Haar measures, it is enough to show that $p_{[0,1)} \circ f^{-1}$ is a Haar measure, i.e. that $p_{[0,1)} \circ f^{-1}$ is invariant with respect to the group operation in S^1, which is the multiplication of complex numbers. Since the exponential function f turns addition of real numbers into multiplication of complex numbers, for $B \subset S^1$ and $z \in \mathbb{C}$ we have

$$f^{-1}(B \cdot z) = f[B] + t \bmod 1 \tag{31}$$

where the translation

$$Y \mapsto Y + t \bmod 1 \tag{32}$$

is the standard shift of set $Y \subset [0,1)$ by t followed by "pulling back" into $[0,1)$ the part of Y that is shifted out of the bounds of $[0,1)$; formally:

$$Y + t \bmod 1 = \big(Y \cap [0, 1-t) \;+t\big) \cup \big(Y \cap [1-t, 1) \;-(1-t)\big)$$

$p_{[0,1)}$ is translation invariant on $[0,1)$ in the sense that for any measurable set $A \subseteq [0,1)$ and $0 \le t < 1$ we have

$$p_{[0,1)}(A) = p_{[0,1)}(A + t \bmod 1),$$

so we have

$$p_H(B \cdot z) = p_{[0,1)}(f^{-1}(B \cdot z)) = p_{[0,1)}(f^{-1}(B) + t \bmod 1) = p_{[0,1)}(f^{-1}(B)) = p_H(B)$$

The Lebesgue measure $p^n_{[0,1)}$ on the n-dimensional cube $[0,1)^n$ also can be regarded as a Haar measure: one can consider the Haar measure p^n_H on the n-dimensional torus

$$T^n = S^1 \times S^1 \times \cdots \times S^1 \;(n \text{ times})$$

which is a compact topological subgroup of $\mathbb{C}^n$ with the coordinate-wise multiplication of complex numbers as group operation. Put

$$f : [0,1)^n \to T^n, \qquad f(t_0, \ldots, t_n) = \big(e^{2\pi i t_0}, \ldots, e^{2\pi i t_n}\big)$$

Then f is a continuous and continuously invertible bijection and, applying the previous argument in each coordinates, one concludes

$$p^n_H = p^n_{[0,1)} \circ f^{-1}$$

References

[1] J. Aaronson. *An Introduction to Infinite Ergodic Theory*, volume 50 of *Mathematical Surveys and Monographs*. American Mathematical Society, Rhode Island, 1997.

[2] S. Bangu. On Bertrand's Paradox. *Analysis*, 70:30–35, 2010.

[3] J.L.F. Bertrand. *Calcul de Probabilités*. Gauthier-Vilars, Paris, 1888.

[4] P. Billingsley. *Ergodic Theory and Information*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London Sydney, 1965.

[5] P. Castell. A consistent restriction of the Principle of Indifference. *The British Journal for the Philosophy of Science*, 49:387–395, 1998.

[6] A. Deitmar and S. Echterhoff. *Principles of Harmonic Analysis*. Universitext. Springer, New York, 2009.

[7] J. Doob. The development of rigor in mathematical probability theory (1900-1950). *American Mathematical Monthly*, pages 586–595, 1996.

[8] H. Fischer. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer, New Yor, Dordrecht, Heidelberg, London, 2011.

[9] A. Hájek. Interpretations of probability. The Stanford Encyclopedia of Philosophy (Summer 2012 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/, 2012. accessed May 29, 2012.

[10] P. Halmos. *Measure Theory*. D. Van Nostrand, New York, 1950.

[11] P. Harremoës. Maximum entropy on compact groups. *Entropy*, 11:222–237, 2009.

[12] E. Jaynes. The Well Posed Problem. *Foundations of Physics*, 4:477–492, 1973.

[13] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation: Foundations of the Theory of Probability, (Chelsea, New York, 1956).

[14] D. Lewis. A subjectivist's guide to objective chance. In *Philosophical Papers, vol. II*, pages 83–132. Oxford University Press, Oxford, 1986.

[15] M. Loéve. *Probability Theory*. D. Van Nostrand, Princeton, Toronto, London, Melbourne, 3rd edition, 1963.

[16] L. Marinoff. A resolution of Bertrand's Paradox. *Philosophy of Science*, 61:1–24, 1994.

[17] J.M. Mikkelson. A resolution of the wine/water paradox. *The British Journal for the Philosophy of Science*, 55:137–145, 2004.

[18] L. Nachbin. *The Haar Integral*. D. Van Nostrand, Princeton, NJ, 1965.

[19] J.S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, Singapore, 2006.

[20] D.W. Rowbottom and N. Schackel. Bangu's random thoughts on Bertrand's Paradox. *Analysis*, 70:689–692, 2010.

[21] W. Rudin. Autohomeomorphisms of compact groups. *Topology and its Applications*, 52:69–70, 1993.

[22] N. Shackel. Bertrand's Paradox and the Principle of Indifference. *Philosophy of Science*, 74:150–175, 2007.

[23] L.E. Szabó. Objective probablity-like things with and without objective indeterminism. *Studies in the History and Philosophy of Modern Physics*, 38:626–634, 2007.

[24] E. K. van Douwen. A compact space with a measure that knows which sets are homeomorphic. *Advances in Mathematics*, 52:1–33, 1984.

[25] R. von Mises. *Probability, Statistics and Truth*. Dover Publications, New York, 2nd edition, 1981. Originally published as 'Wahrscheinlichkeit, Statistik und Wahrheit' (Springer, 1928).

[26] J. von Plato. *Creating Modern Probability*. Cambridge Studies in Probability , Induction and Decision Theory. Cambridge University Press, Cambridge, 1994.

[27] A. Wightman. Hilbert's 6th problem. In F.E Browder, editor, *Mathematical Developments Arising from Hilbert Problems: Proceedings*, volume 28 of *Proceedings of Symposia in Pure Mathematics*, pages 147–240. American Mathematical Society, 1983.

Bell inequality and common causal explanation in algebraic quantum field theory

Gábor Hofer-Szabó[*]
Péter Vecsernyés[†]

Abstract

Bell inequalities, understood as constraints between classical conditional probabilities, can be derived from a set of assumptions representing a common causal explanation of classical correlations. A similar derivation, however, is not known for Bell inequalities in algebraic quantum field theories establishing constraints for the expectation of specific linear combinations of projections in a quantum state. In the paper we address the question as to whether a 'common causal justification' of these non-classical Bell inequalities is possible. We will show that although the classical notion of common causal explanation can readily be generalized for the non-classical case, the Bell inequalities used in quantum theories cannot be derived from these non-classical common causes. Just the opposite is true: for a set of correlations there can be given a non-classical common causal explanation *even if* they violate the Bell inequalities. This shows that the range of common causal explanations in the non-classical case is wider than that restricted by the Bell inequalities.

Key words: Bell inequality, common cause, noncommutativity, algebraic quantum field theory.

1 Introduction

The original context which led to the formulation of the Bell inequalities was the intention to accomodate quantum correlations in a *locally causal* theory. The clearest formulation of such a theory is due to Bell himself (Bell, 1987, p. 54). In a number of seminal papers Bell carefully analyzed the intuitions lying behind our notion of locality and causality. His major contribution, however, consisted in translating these intricate notions into a simple probabilistic language which made these notions tractable both for mathematical treatment and later for experimental testability. This probabilistic framework made it possible to exactly identify the probabilistic requirements responsible for the violation of the Bell inequalities in the EPR scenario. A decade later authors like Van Fraassen (1982), Jarrett (1984) and Shimony (1986) spent much time to analyze the philosophical consequences of giving up either the one or the other of these probabilistic assumptions. It also turned out soon that the conceptual framework in which the Bell inequalities can be treated most naturally is the common causal explanation of correlations, originally stemming from Reichenbach (1956) and later adopted to the EPR case by Van Fraassen (1982).

Since the aim of these considerations was to accomodate the EPR scenario in a classical world picture, both Bell and the subsequent writers used a *classical* probabilistic framework in their analysis. All the assumptions representing locality and causality and also the resulting Bell inequalities were formulated in the language of the classical probability theory. Now, if the Bell inequalities were classical, how could they be violated in the EPR scenario which is well known to be described by

[*]King Sigismund College, Budapest, email: gsz@szig.hu
[†]Wigner Research Centre for Physics, Budapest, email: vecsernyes.peter@wigner.mta.hu

quantum theory? Well, the answer is that quantum theory with its mathematical structure and ontological commitments played *no role at all* in the Bell scenario. Quantum mechanics was only used *to generate classical probabilities*, more specifically, *classical conditional probabilities* by the Born rule. These classical conditional probabilities, however, could also have been gained directly from the experiments, and indeed later they have been gained so. In other words, the original context of the Bell inequalities has no intimate link to quantum theory even if quantum theory produces probabilities which, reinterpreted as classical conditional probabilities, violate those inequalities. This classical view on the Bell inequalities manifests itself in various authors. Nicolas Gisin for example writes: "Bell inequalities are relations between conditional probabilities valid under the locality assumption." (Gisin 2009, p. 126)

In the face of all these, the Bell inequality has made its way into quantum theory. It has been soon formulated as a general mark of entanglement of the given quantum state on a C^*-algebra (Summers and Werner 1987a, b). A quote from Bengtson and Zyczkowski (2006, p. 362) might illustrate this change of focus in the role of Bell inequalities: "The Bell inequalities may be viewed as a kind of separability criterion, related to a particular entanglement witness, so evidence of their violation for certain states might be regarded as an experimental detection of quantum entanglement." How could the Bell inequality make its way to this non-classical formalism so alien from its original context? Does there exist a justification for this 'trespass'?

In this paper we would like to investigate a possible justification for this transition. In this justification we intend to follow the route pioneered by Bell, Van Fraassen, Jarrett, Shimony and others in that we stick to the conviction that the Bell inequalities follow from the requirement of implementing correlations into a *locally causal* theory. We transcend, however, this view in *not* assuming that this theory has to be *classical*. Or in other words, we pose the question whether the probabilistic requirements representing local causality and constituting the core of the Bell inequalities can be reasonable formulated also in a non-classical theory.

A natural candidate for such a non-classical theory with clear conceptions of locality and causality is algebraic quantum field theory (AQFT) (Haag, 1992). In AQFT events are represented by projections with well defined spacetime support and local causality is ensured by a set of axioms. Hence we can pose the question as to whether the Bell inequalities featuring in AQFT follow from a locally causal explanation of correlations in a similar manner to the classical case. Since we intend to give a causal explanation for *correlations between events*, therefore causal explanation is meant to be a *common causal* explanation. We will see that the connection between a common causal explanation and the Bell inequalities in AQFT is not so tight as in the classical case. In the classical case common causes necessarily commute (in the set theoretical 'meet' operation) with their effects. Since the quantum events of AQFT form a noncommutative structure, one can decide whether to require that common causes commute with their effects or not. If commutativity is required, the Bell inequalities will follow from the common cause just like in the classical case. But, as we will argue, requiring commutativity is only a remininscence of the classical treatment of correlations and is completely unjustified in the quantum case (see e.g. (Clifton, Ruetsche 1999)). For noncommuting common causes the Bell inequalities will turn out *not* to be derivable from the presence of the common cause—at least not in the similar way to the the classical derivation. This raises the question whether correlations violating the Bell inequalities can have a noncommuting common causal explanation. We will answer this question in the affirmative showing up a situation when a set of correlations maximally violating a specific type Bell inequality has a common causal explanation, which is local in the sense that it can be accomodated in the intersection of the causal pasts of the correlating events. The model we use for this example is the local quantum Ising model, the simplest AQFT with locally finite degrees of freedom.

The paper is structured as follows. In Section 2 we briefly collect the most important concepts and some of the representative propositions concerning the Bell inequality in AQFT. In Section 3 and 4 we give the definition of the classical and the non-classical common causal explanations, respectively,

and show how these explanations relate to the Bell inequalities. Since the correct 'translation' of the so-called locality and no-conspiracy conditions of the classical common causal explanation into the non-classical setting is a subtle point not needed for our main purpose, we transfer it into the Appendix. Now, the common causal explanations in the EPR-Bell scenario is always meant as providing a *joint* common cause for a *set* of correlations. Providing a joint common cause for a set of correlations is much more demanding than simply providing a common cause for a *single* correlation. Therefore in Section 5, preparing for the more complicated case, we investigate the possibility of a common causal explanation of a single correlation, or in the philosophers' jargon, the status of the Common Causal Principle in AQFT. In Section 6 we return to our original question and present a noncommutative common causal explanation for a *set* of correlations maximally violating some Bell inequalities. In Section 7 we briefly analyze the philosophical consequences of applying *noncommuting* common causes in our causal explanation. We conclude the paper in Section 8.

2 The Bell inequality in algebraic quantum field theory

In this Section we collect the most important concepts and some of the representative propositions concerning the Bell inequality in AQFT (see (Summers 1990) and (Halvorson 2007)). We start with the general C^*-algebraic setting and then go over to the special algebraic quantum field theoretical formulation.

In the general C^*-algebraic setting Bell inequality is treated in the following way. Let $\mathcal{A}$ and $\mathcal{B}$ be two mutually commuting C^*-subalgebras of some C^*-algebra $\mathcal{C}$. A *Bell operator* R for the pair $(\mathcal{A},\mathcal{B})$ is an element of the following set:

$$\mathbb{B}(\mathcal{A},\mathcal{B}) \quad := \quad \left\{ \frac{1}{2}\big(X_1(Y_1+Y_2)+X_2(Y_1-Y_2)\big) \,\big|\, X_i = X_i^* \in \mathcal{A};\, Y_i = Y_i^* \in \mathcal{B};\, -\mathbf{1} \leqslant X_i, Y_i \leqslant \mathbf{1} \right\}$$

where $\mathbf{1}$ is the unit element of $\mathcal{C}$. For any Bell operator R the following can be proven:

Theorem 1. For any state $\phi\colon \mathcal{C} \to \mathbf{C}$, one has $|\phi(R)| \leqslant \sqrt{2}$.

Theorem 2. For separable states (i.e. for convex combinations of product states) $|\phi(R)| \leqslant 1$.

The *Bell correlation coefficient* of a state ϕ is defined as

$$\beta(\phi,\mathcal{A},\mathcal{B}) \quad := \quad \sup\big\{ |\phi(R)| \,\big|\, R \in \mathbb{B}(\mathcal{A},\mathcal{B}) \big\}$$

and the *Bell inequality* is said to be *violated* if $\beta(\phi,\mathcal{A},\mathcal{B}) > 1$, and *maximally violated* if $\beta(\phi,\mathcal{A},\mathcal{B}) = \sqrt{2}$. An important result of Bacciagaluppi (1994) is the following:

Theorem 3. If $\mathcal{A}$ and $\mathcal{B}$ are C^*-algebras, then there are some states violating the Bell inequality for $\mathcal{A}\otimes\mathcal{B}$ iff both $\mathcal{A}$ and $\mathcal{B}$ are non-abelian.

Going over to von Neumann algebras Landau (1987) has shown that the maximal violation of the Bell inequality is generic in the following sense:

Theorem 4. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be von Neumann algebras, and suppose that $\mathcal{N}_1$ is abelian and $\mathcal{N}_1 \subseteq \mathcal{N}_2'$ ($\mathcal{N}'$ being the commutant of $\mathcal{N}$). Then for any state $\beta(\phi,\mathcal{A},\mathcal{B}) \leqslant 1$. On the other hand, if both $\mathcal{N}_1$ and $\mathcal{N}_2$ are non-abelian von Neumann algebras such that $\mathcal{N}_1 \subseteq \mathcal{N}_2'$, and if $(\mathcal{N}_1,\mathcal{N}_2)$ satisfies the *Schlieder-property*,[1] then there is a state ϕ for which $\beta(\phi,\mathcal{A},\mathcal{B}) = \sqrt{2}$.

[1] The commuting pair $(\mathcal{A},\mathcal{B})$ of C^*-subalgebras in $\mathcal{C}$ obeys the Schlieder-property, if for $0 \neq A \in \mathcal{A}$ and $0 \neq B \in \mathcal{B}$, $AB \neq 0$. Since in case of von Neumann algebras A and B can be required to be projections, Schlieder-property is the analogue of logical independence in classical logic.

Adding further constraints on the von Neumann algebras one obtains other important results such as the following two:

Theorem 5. If $\mathcal{N}_1$ and $\mathcal{N}_2$ are *properly infinite*[2] von Neumann algebras on the Hilbert space $\mathcal{H}$ such that $\mathcal{N}_1 \subseteq \mathcal{N}_2'$, and $(\mathcal{N}_1, \mathcal{N}_2)$ satisfies the Schlieder-property, then there is a dense set of vectors in $\mathcal{H}$ inducing states which violate the Bell inequality across $(\mathcal{N}_1, \mathcal{N}_2)$ (Halvorson and Clifton, 2000).

Theorem 6. Let $\mathcal{H}$ be a separable Hilbert space and let $\mathcal{R}$ be a von Neumann factor of type III_1 acting on $\mathcal{H}$. Then every normal state ϕ of $\mathcal{B}(\mathcal{H})$ maximally violates the Bell inequality across $(\mathcal{R}, \mathcal{R}')$ (Summers and Werner, 1988).

Type III factors featuring in Theorems 5-6. are the typical local von Neumann algebras in AQFT with locally infinite degrees of freedom. Here we briefly survey the basic notions of the theory.

In AQFT observables (including quantum events) are represented by unital C^*-algebras associated to bounded regions of a given spacetime. The association of algebras and spacetime regions is established along the following lines.

(i) *Isotony.* Let $\mathcal{S}$ be a spacetime. A *double cone* in $\mathcal{S}$ is the intersection of the causal past of a point x with the causal future of a point y timelike to x. Let $\mathcal{K}$ be a collection of double cones of $\mathcal{S}$ 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 observable 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.

(ii) *Microcausality.* The net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ satisfies microcausality (aka Einstein causality): $\mathcal{A}(V')' \cap \mathcal{A} \supseteq \mathcal{A}(V), V \in \mathcal{K}$, where primes denote spacelike complement and algebra commutant, respectively. $\mathcal{A}(V')$ is the smallest C^*-algebra in $\mathcal{A}$ containing the local algebras $\mathcal{A}(\tilde{V}), \mathcal{K} \ni \tilde{V} \subset V'$.

(iii) *Covariance.* Let $\mathcal{P}_\mathcal{K}$ be the subgroup of the group $\mathcal{P}$ of geometric symmetries of $\mathcal{S}$ 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: $\alpha_g(\mathcal{A}(V)) = \mathcal{A}(g \cdot V), V \in \mathcal{K}$.

To the net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ satisfying the above requirements we will refer to as a $\mathcal{P}_\mathcal{K}$-*covariant local quantum theory.* If $\mathcal{S} = \mathcal{M}$ is the Minkowski spacetime and $\mathcal{K}$ is the net of all double cones then $\mathcal{P}_\mathcal{K}$ is the Poincaré group, and we obtain Poincaré covariant algebraic quantum field theories with locally infinite degrees of freedom. Restricting the collection $\mathcal{K}$ one can obtain $\mathcal{P}_\mathcal{K}$-covariant local quantum theories with locally finite degrees of freedom, for instance our example, the local quantum Ising model (see below).

A *state* ϕ in a local quantum 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 expectation value of a projection is the probability of the event that the observable takes on the value 1 in the appropriate quantum state. Two commuting quantum events A and B are said to be *correlating* in a state ϕ if

$$\phi(AB) \quad \neq \quad \phi(A)\phi(B).$$

[2] The center contains no finite projections.

If the events are supported in spatially separated spacetime regions V_A and V_B, respectively, then the correlation between them is said to be *superluminal*. To see that superluminal correlations violating Bell inequalities abound in Poincaré covariant algebraic quantum field theories, one has to introduce further requirements on the representations of $\mathcal{A}$ (see Haag 1992):

(iv) *Unitary implementability.* There is a strongly continuous unitary representation of the Poincaré group, $U\colon \mathcal{P} \to \mathcal{B}(\mathcal{H}_\phi)$, such that

$$\pi_\phi(\alpha_g(A)) = U(g)\pi_\phi(A)U(g)^*, \qquad A \in \mathcal{A},\ g \in \mathcal{P}.$$

(v) *Vacuum condition.* There is a (up to a scalar) unique vector Ω in the Hilbert space $\mathcal{H}_0$ corresponding to the vacuum state ϕ_0 such that $U(g)\Omega = \Omega$ for all $g \in \mathcal{P}$.

(vi) *Spectrum condition.* The spectrum of the self-adjoint generators of the strongly continuous unitary representation of the translation subgroup $\mathbf{R}^4$ of $\mathcal{P}$ lies in the closed forward light cone.

(vii) *Weak additivity.* For any nonempty open region V, the set of operators $\cup_{g\in\mathbb{R}^4}\mathcal{N}(g\cdot V)$ is dense in $\mathcal{B}(\mathcal{H}_0)$ (in the weak operator topology).

Now, under conditions (i)-(vii) the local von Neumann algebras supported in spacelike separated double cones satisfy the Schlieder property (Schlieder, 1969). Therefore Theorem 4 applies to these algebras stating that there is a state maximally violating the Bell inequality across these local algebras. Moreover, if the net is *non-trivial*[3], then the local von Neumann algebras are properly infinite. This makes Theorem 5 applicable to local von Neumann algebras supported in spacelike separated double cones stating that there is a dense set of vectors in $\mathcal{H}$ inducing states which violate the Bell inequality.

Being properly infinite the von Neumann algebras cannot be of type I_n and II_1 but they still can be of type I_∞ or II_∞ . However, a set of independent results indicates that the local von Neumann algebras are of type III, more specifically *hyperfinite*[4] factors of type III_1. Buchholz et al. (1987) proved that the local algebras for relativistic free fields are type III_1 and it was also shown that one can construct the local von Neumann algebras as a unique type III_1 hyperfinite factor from the underlying Wightman theory by adding the assumption of *scaling limit* (see (Fredenhagen (1985)).

Instead of deriving the type of the von Neumann algebras from more general physical requirements, one also can explicitely add this condition as a new axiom of AQFT:

(viii) *The type of the algebras.* For every double cone V the von Neumann algebra $\mathcal{N}(V)$ is of type III_1.

Under conditions (i)-(viii) the local von Neumann algebras supported in spacelike separeted double cones satisfy the assumptions of Theorem 6, therefore every normal state will maximally violate the Bell inequality across pairs of algebras supported in spacelike separated double cones.

Finally, we mention a physically important consequence of Theorem 6:

Theorem 7. The vacuum state maximally violates the Bell inequality across the *wedge*[5] algebras $(\mathcal{N}(W),\mathcal{N}(W)')$. (Summers, Werner 1988).

As said above, the Bell inequality typically used in AQFT is of the following form:

$$\left|\phi\big(X_1(Y_1+Y_2)+X_1(Y_1-Y_2)\big)\right| \leqslant 2, \tag{1}$$

[3]For each double cone V, $\mathcal{A}(V) \neq \mathbb{C}\mathbf{1}$.
[4]The weak closure of an ascending sequence of finite dimensional algebras.
[5]Poincaré transforms of the region $W_R := \{x \in \mathcal{M} | x_1 > |x_0|\}$.

where $X_m \in \mathcal{N}(V_A)$ and $Y_n \in \mathcal{N}(V_B)$ are self-adjoint *contractions* (that is $-\mathbf{1} \leqslant X_m, Y_n \leqslant \mathbf{1}$ for $m, n = 1, 2$) supported in spatially separated spacetime regions V_A and V_B, respectively. This type of Bell inequality is usually referred to as the *Clauser–Horne–Shimony–Holte (CHSH) inequality* (Clauser, Horne, Shimony and Holt, 1969). Sometimes in the EPR-Bell literature another Bell-type inequality is used instead of (1): the *Clauser–Horne (CH) inequality* (Clauser and Horne, 1974) defined in the following way:

$$-1 \leqslant \phi(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1) \leqslant 0, \tag{2}$$

where A_m and B_n are *projections* located in $\mathcal{N}(V_A)$ and $\mathcal{N}(V_B)$, respectively. It is easy to see, however, that the two inequalities are equivalent: in a given state ϕ the set $\{(A_m, B_n); m, n = 1, 2\}$ violates the CH inequality (2) *if and only if* the set $\{(X_m, Y_n); m, n = 1, 2\}$ of self-adjoint contractions given by

$$X_m := 2A_m - \mathbf{1} \tag{3}$$
$$Y_n := 2B_n - \mathbf{1} \tag{4}$$

violates the CHSH inequality (1). Therefore, from now on we will concentrate only on the CH-type Bell inequalities.

In the next two sections we turn to the common causal explanation behind the Bell inequalities. In the next Section we introduce the basic notions of the classical common causal explanation leading to the Bell inequalities; in the subsequent Section we generalize these notions for the quantum case.

3 Classical common causal explanation

Let us begin with Hans Reichenbach's (1956) original definition which is historically the first probabilistic characterization of the notion of the common cause. Let (Ω, Σ, p) be a classical probability measure space and let A and B be two positively correlating events in Σ:

$$p(A \wedge B) > p(A)\, p(B). \tag{5}$$

Definition 1. An event $C \in \Sigma$ is said to be the *Reichenbachian common cause* of the correlation between events A and B if the following conditions hold:

$$p(A \wedge B|C) = p(A|C)p(B|C) \tag{6}$$
$$p(A \wedge B|C^{\perp}) = p(A|C^{\perp})p(B|C^{\perp}) \tag{7}$$
$$p(A|C) > p(A|C^{\perp}) \tag{8}$$
$$p(B|C) > p(B|C^{\perp}) \tag{9}$$

where $C^{\perp}$ denotes the orthocomplement of C and $p(\cdot\,|\,\cdot)$ is the conditional probability defined by the Bayes rule. One refers to equations (6)-(7) as the *screening-off conditions* and to inequalities (8)-(9) as the *positive statistical relevancy conditions*.

Reichenbach's definition, however, cannot be applied directly to AQFT for four reasons. First, the positive statistical relevancy conditions restrict one to common causes which increase the probability of their effects; or in other words, they exclude negative causes. Second, the definition also excludes situations in which the correlation is *not* due to a *single* cause but to a *system* of cooperating common causes. Third, it is silent about the spatiotemporal localization of the events. Fourth and most importantly, it is classical.

Let us first address the first two problems. Let A and B be two correlating events in a classical probability measure space (Ω, Σ, p) that is

$$p(A \wedge B) \neq p(A)\, p(B). \tag{10}$$

Definition 2. A partition $\{C_k\}_{k\in K}$ in Σ is said to be the *common cause system* of the correlation (10) if the following screening-off condition holds for all $k \in K$:

$$p(A \wedge B|C_k) \quad = \quad p(A|C_k)\, p(B|C_k), \tag{11}$$

where $|K|$, the cardinality of K is said to be the *size* of the common cause system. A common cause system of size 2 is called a common cause (without the adjective 'Reichenbachian', indicating that the inequalities (8)-(9) are not required).

Concerning the third problem, namely, the localization of the common cause, one has (at least) three different options. Suppose that the two events A and B are localized in two bounded and spatially separated regions V_A and V_B of a spacetime $\mathcal{S}$. Then one can localize $\{C_k\}$ either (i) in the *union* or (ii) in the *intersection* of the causal past of the regions V_A and V_B; or (iii) more restrictively, in the spacetime region which lies in the intersection of causal pasts of *every* point of $V_A \cup V_B$. Formally, we have

$$\begin{aligned} wpast(V_A, V_B) &:= I_-(V_A) \cup I_-(V_B) \\ cpast(V_A, V_B) &:= I_-(V_A) \cap I_-(V_B) \\ spast(V_A, V_B) &:= \cap_{x\in V_A\cup V_B}\, I_-(x) \end{aligned}$$

where $I_-(V)$ denotes the union of the backward light cones i.e. the causal pasts $I_-(x)$ of every point x in V (Rédei, Summers 2007). We will refer to the above three pasts in turn as the *weak past*, *common past*, and *strong past* of A and B, respectively (see Fig. 1). The notion of these pasts presupposes a spacetime localization structure of the classical event algebra. (For such an attempt see (Henson, 2005).)

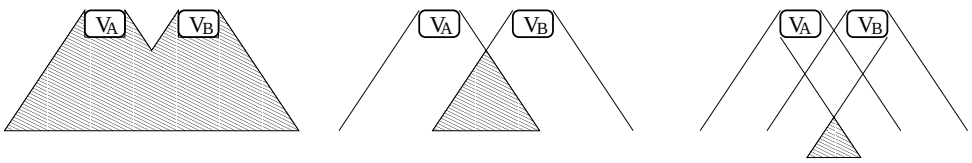


Figure 1: Possible localizations of the common cause system in different pasts of V_A and V_B.

Now, suppose that we do not face *one* correlation (A, B) but a *set* of correlations that is events A_m and B_n in Σ such that for any $m \in M, n \in N$

$$p(A_m \wedge B_n) \quad \neq \quad p(A_m)\, p(B_n). \tag{12}$$

If our aim is to explain all of these pair-correlations $\{(A_m, B_n); m \in M, n \in N\}$ by a *single* common cause system, then we are led to the following definition:

Definition 3. A partition $\{C_k\}_{k\in K}$ in Σ is said to be a *joint*[6] *common cause system* of the set of correlations $\{(A_m, B_n); m \in M, n \in N\}$ if the following screening-off condition holds for all $m \in M$, $n \in N$, and $k \in K$:

$$p(A_m \wedge B_n|C_k) \quad = \quad p(A_m|C_k)\, p(B_n|C_k). \tag{13}$$

[6] In (Hofer-Szabó and Vecsernyés, 2012a,b) called *common* common cause system.

Obviously, for a set of correlations to have a joint common cause system is much more demanding than to simply have a *separate* common cause system for each correlation.

Now, let us complicate the picture a little further by introducing *conditional* probabilities. Suppose that events A_m and B_n are *outcomes* of measurements of the observables $\mathrm{A_m}$ and $\mathrm{B_n}$, respectively. Let a_m and b_n, respectively denote the events that the appropriate measurement devices are set to measure the observables $\mathrm{A_m}$ and $\mathrm{B_n}$, respectively. Let us refer to these events as *measurement choices*. To be more specific, suppose that each measurement choice a_m in region V_A can yield only two outcomes A_m and $A_m^{\perp}$, and similarly the measurement choices b_n in region V_B can again yield only two outcomes B_n and $B_n^{\perp}$. Finally, suppose that probability of the different measurement choices a_m in region V_A add up to 1, and similarly for the measurement choices b_n in region V_B.

Now, the events A_m and B_n are said to be correlating in the *conditional* sense if for all A_m, B_n, a_m, $b_n \in \Sigma$ $(m \in M, n \in N)$ the following holds:

$$p(A_m \wedge B_n \,|\, a_m \wedge b_n) \quad \neq \quad p(A_m|a_m \wedge b_n)\, p(B_n|a_m \wedge b_n). \tag{14}$$

What does a joint common causal explanation of these conditional correlations consists in? The answer to this question is given in the following definition:

Definition 4. A *local, non-conspiratorial joint common causal explanation* of the conditional correlations (14) consists in providing a partition $\{C_k\}$ in Σ such that for any $m, m' \in M, n, n' \in N$ the following requirements hold:

$$p(A_m \wedge B_n|a_m \wedge b_n \wedge C_k) = p(A_m|a_m \wedge b_n \wedge C_k)\, p(B_n|a_m \wedge b_n \wedge C_k) \qquad \text{(screening-off)} \tag{15}$$

$$p(A_m|a_m \wedge b_n \wedge C_k) = p(A_m|a_m \wedge b_{n'} \wedge C_k) \qquad \text{(locality)} \tag{16}$$

$$p(B_n|a_m \wedge b_n \wedge C_k) = p(B_n|a_{m'} \wedge b_n \wedge C_k) \qquad \text{(locality)} \tag{17}$$

$$p(a_m \wedge b_n \wedge C_k) = p(a_m \wedge b_n)\, p(C_k) \qquad \text{(no-conspiracy)} \tag{18}$$

The motivation behind requirements (15)-(18) is the following. *Screening-off* (15) is simply the application of the notion of common cause for conditional correlations: although A_m and B_n are correlating conditioned on a_m and b_n, they will cease to do so if we further condition on $\{C_k\}$. *Locality* (16)-(17) is the natural requirement that the measurement outcome on the one side should depend only on the measurement choice on the same side and the value of the common cause but not on the measurement choice on the opposite side. Finally, no-conspiracy (18) is the requirement that the common cause system and the measurement choices should be probabilistically independent. (For the justification of the above requirements by Causal Markov Condition see (Glymour, 2006).)

Let us now proceed further. A straightforward consequence of Definition 4 is the following proposition (Clauser, Horne, 1974):

Proposition 1. Let A_m, B_n, a_m and b_n $(m, n = 1, 2)$ be eight events in a classical probability measure space (Ω, Σ, p) such that the pairs $\{(A_m, B_n); m, n = 1, 2\}$ correlate in the conditional sense of (14). Suppose that $\{(A_m, B_n); m, n = 1, 2\}$ has a local, non-conspriratorial joint common causal explanation in the sense of Definition 4. Then for any $m, m', n, n' = 1, 2; m \neq m'; n \neq n'$ the following *classical Clauser–Horne* inequality holds:

$$\begin{aligned} -1 \leqslant p(A_m \wedge B_n|a_m \wedge b_n) + p(A_m \wedge B_{n'}|a_m \wedge b_{n'}) + p(A_{m'} \wedge B_n|a_{m'} \wedge b_n) \\ -p(A_{m'} \wedge B_{n'}|a_{m'} \wedge b_{n'}) - p(A_m|a_m \wedge b_n) - p(B_n|a_m \wedge b_n) \leqslant 0 \end{aligned} \tag{19}$$

Proof. It is an elementary fact of arithmetic that for any $\alpha, \alpha', \beta, \beta' \in [0, 1]$ the number

$$\alpha\beta + \alpha\beta' + \alpha'\beta - \alpha'\beta' - \alpha - \beta \tag{20}$$

lies in the interval $[-1, 0]$. Now let $\alpha, \alpha', \beta, \beta'$ be the following conditional probabilities:

$$\begin{aligned} \alpha &:= p(A_m|a_m \wedge b_n \wedge C_k) && (21)\\ \alpha' &:= p(A_{m'}|a_{m'} \wedge b_{n'} \wedge C_k) && (22)\\ \beta &:= p(B_n|a_m \wedge b_n \wedge C_k) && (23)\\ \beta' &:= p(B_{n'}|a_{m'} \wedge b_{n'} \wedge C_k) && (24) \end{aligned}$$

Plugging (21)-(24) into (20) and using locality (16)-(17) one obtains

$$\begin{aligned} -1 \leqslant p(A_m|a_m \wedge b_n \wedge C_k)p(B_n|a_m \wedge b_n \wedge C_k) + p(A_m|a_m \wedge b_{n'} \wedge C_k)p(B_{n'}|a_m \wedge b_{n'} \wedge C_k) \\ +p(A_{m'}|a_{m'} \wedge b_n \wedge C_k)p(B_n|a_{m'} \wedge b_n \wedge C_k) - p(A_{m'}|a_{m'} \wedge b_{n'} \wedge C_k)p(B_{n'}|a_{m'} \wedge b_{n'} \wedge C_k) \\ -p(A_m|a_m \wedge b_n \wedge C_k) - p(B_n|a_m \wedge b_n \wedge C_k) \leqslant 0 \end{aligned} \quad (25)$$

Using screening-off (15) one gets

$$\begin{aligned} -1 \leqslant p(A_m \wedge B_n|a_m \wedge b_n \wedge C_k) + p(A_m \wedge B_{n'}|a_m \wedge b_{n'} \wedge C_k) + p(A_{m'} \wedge B_n|a_{m'} \wedge b_n \wedge C_k) \\ -p(A_{m'} \wedge B_{n'}|a_{m'} \wedge b_{n'} \wedge C_k) - p(A_m|a_m \wedge b_n \wedge C_k) - p(B_n|a_m \wedge b_n \wedge C_k) \leqslant 0 \end{aligned} \quad (26)$$

Multiplying the above inequality by $p(C_k)$, using no-conspiracy (18) and summing up for the index k one obtains

$$\begin{aligned} -1 \leqslant \sum_k \big(p(A_m \wedge B_n \wedge C_k|a_m \wedge b_n) + p(A_m \wedge B_{n'} \wedge C_k|a_m \wedge b_{n'}) + p(A_{m'} \wedge B_n \wedge C_k|a_{m'} \wedge b_n) \\ -p(A_{m'} \wedge B_{n'} \wedge C_k|a_{m'} \wedge b_{n'}) - p(A_m \wedge C_k|a_m \wedge b_n) - p(B_n \wedge C_k|a_m \wedge b_n)\big) \leqslant 0 \end{aligned} \quad (27)$$

Finally, applying the theorem of total probability

$$\sum_k p(Y \wedge C_k) = p(Y)$$

one arrives at (19) which completes the proof. ■

Proposition 1 plays a crucial role in understanding the CH inequality (19). It provides, so to say, a 'classical common causal justification' of the classical CH inequality by showing that (19) is a necessary condition for the existence of a local, non-conspriratorial joint common causal explanation for a set of conditional correlations.

The well-known situation in which the classical CH inequality (19) is violated and hence the correlations in question have no local, non-conspriratorial joint common causal explanation, is the EPR-Bohm scenario. Consider a pair of spin-$\frac{1}{2}$ particles prepared in the singlet state (see Fig. 2). Let a_m $(m = 1, 2)$ denote the event that the measurement apparatus is set to measure the spin

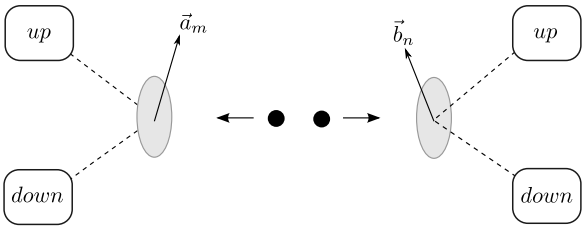


Figure 2: EPR–Bohm setup for spin-$\frac{1}{2}$ particles

in direction $\vec{a}_m$ in the left wing; and let $p(a_m)$ stand for the probability of a_m. Let b_n $(n = 1, 2)$ and $p(b_n)$ respectively denote the same for direction $\vec{b}_n$ in the right wing. (Note that $m = n$ does not mean that $\vec{a}_m$ and $\vec{b}_n$ are parallel directions.) Furthermore, let $p(A_m)$ stand for the probability that the spin measurement in direction $\vec{a}_m$ in the left wing yields the result 'up' and let $p(B_n)$ be defined in a similar way in the right wing for direction $\vec{b}_n$. According to the statistical algorithm of quantum mechanics the conditional probability of getting an 'up' result *provided* we measure the spin in direction $\vec{a}_m$ in the left wing; getting an 'up' result *provided* we measure the spin in direction $\vec{b}_n$ in the right wing; and getting 'up-up' result *provided* we measure the spin in both directions $\vec{a}_m$ and $\vec{b}_n$ are given by the following relations:

$$p(A_m|a_m \wedge b_n) = \frac{1}{2} \tag{28}$$

$$p(B_n|a_m \wedge b_n) = \frac{1}{2} \tag{29}$$

$$p(A_m \wedge B_n|a_m \wedge b_n) = \frac{1}{2}\sin^2\left(\frac{\theta_{a_m b_n}}{2}\right) \tag{30}$$

where $\theta_{a_m b_n}$ denotes the angle between directions $\vec{a}_m$ and $\vec{b}_n$. For non-perpendicular directions $\vec{a}_m$ and $\vec{b}_n$ (28)-(30) predict conditional correlations specified in (14). Now, in order to provide a *classical* local, non-conspiratorial joint common causal explanation for these correlations, the conditional probabilities (28)-(30) have to satisfy the classical CH inequality (19). Since for appropriate choice of the measurement directions this inequalitity is violated, EPR correlations cannot be given a *classical* local, non-conspiratorial joint common causal explanation.

Observe that up to this point everything has been classical. Quantum mechanics (QM) was simply used to generate classical conditional probabilities by the Born rule. These conditional probabilities, however, could also have been directly obtained from the laboratory and in the actual experiments they are gained in this direct way indeed. So it is completely satisfactory to interpret the EPR scenario—in accord with the quote from Gisin in the Introduction—as a classical situation with classical conditional correlation (between detector clicks) violating the classical CH inequality (19) (see (Szabó 1998)).

But this is *not* the standard interpretation. The standard way to describe the above EPR situation is to adopt another mathematical formalism, the formalism of quantum theory. Here events are represented as projections of the von Neumann lattice of the tensor product matrix algebra $M_2(\mathbf{C}) \otimes M_2(\mathbf{C})$ and probabilities are gained by the quantum states. So instead of (28)-(30) one writes the following:

$$\phi^s(A_m) = Tr\big(\rho^s\,(A_m \otimes \mathbf{1}_B)\big) = \frac{1}{2} \tag{31}$$

$$\phi^s(B_n) = Tr\big(\rho^s\,(\mathbf{1}_A \otimes B_n)\big) = \frac{1}{2} \tag{32}$$

$$\phi^s(A_m B_n) = Tr\big(\rho^s\,(A_m \otimes B_n)\big) = \frac{1}{2}\sin^2\left(\frac{\theta_{a_m b_n}}{2}\right) \tag{33}$$

where A_m and B_n denote projections onto the eigensubspaces with eigenvalue $+\frac{1}{2}$ of the spin operators associated with directions $\vec{a}_m$ and $\vec{b}_n$, respectively, and $\phi^s(\,\cdot\,) = Tr(\rho^s\,\cdot\,)$ is the singlet state. Moreover, if we go over to AQFT, these projections will be localized in a well-defined spacetime region.

Substituting the *non-classical* probabilities (31)-(33) into the *non-classical* CH inequality (2) defined in the Introduction one finds a violation of this inequality for appropriate choices of the projections A_m, B_n. But what does it mean? First, it is important to be aware of the fact that now

we adopt another theory to account for correlations. But then we need to take the consequences of this move seriously. This means that we need to represent *every* event of the model as projections of a von Neumann algebra. Among them common causes! So the following questions arise: Can the classical notion of the common cause (system) generalized for the non-classical case? What is the relation of this non-classical notion of common cause to the non-classical CH inequality (2)? Does there exist a non-classical common causal justification of the Bell inequalities used in AQFT similar to the classical one?

As it will turn out soon, one can generalize the notion of the common cause also for the algebraic quantum field theoretical setting, and one can also give a precise definition of a local, non-conspiratorial joint common causal explanation of a set of correlations in AQFT. However, it also will turn out that there is no direct relation between this common causal explanation and the Bell inequalities. Or to put it briefly, correlation violating the Bell inequality can still have a local, non-conspiratorial joint common causal explanation. In order to see all these, first we have to generalize the notions of this Section to the quantum case.

4 Non-classical common causal explanation

Let us first generalize the notion of the common cause system to the quantum case in the following way. Replace the classical probability measure space (Ω, Σ, p) by the non-classical probability measure space $(\mathcal{N}, \mathcal{P}(\mathcal{N}), \phi)$ where $\mathcal{P}(\mathcal{N})$ is the (non-distributive) lattice of projections (events) and ϕ is a state of a von Neumann algebra $\mathcal{N}$. We note that in case of projection lattices we will use only algebra operations (products, linear combinations) instead of lattice operations $(\vee, \wedge)$. In case of commuting projections $A, B \in \mathcal{P}(\mathcal{N})$ lattice operations can be given in terms of algebraic operations.

A set of mutually orthogonal projections $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ is called a *partition of the unit* $\mathbf{1} \in \mathcal{N}$ if $\sum_k C_k = \mathbf{1}$. Two commuting projections A and $B \in \mathcal{P}(\mathcal{N})$ are said to be correlating in the state $\phi\colon \mathcal{N} \to \mathbf{C}$ if

$$\phi(AB) \quad \neq \quad \phi(A)\,\phi(B). \tag{34}$$

Since ϕ is linear, a kind of 'theorem of total probablity', $\sum_i \phi(AP_i) = \phi(A\sum_i P_i) = \phi(A)$, holds for any partition $\{P_i\}$ of the unit, hence (34) is equivalent to

$$\phi(AB)\,\phi(A^\perp B^\perp) \quad \neq \quad \phi(AB^\perp)\,\phi(A^\perp B). \tag{35}$$

Now, following the lines of Definition 2 one can characterize the non-classical common cause system of the correlation (34) as a screener-off partition of the unit. To make the definition meaningful we have to introduce the following *conditional expectation* $E_c\colon \mathcal{N} \to \mathcal{C}$:

$$E_c(A) := \sum_{k\in K} C_k A C_k, \tag{36}$$

where $\{C_k\}_{k\in K}$ is a partition of the unit of $\mathcal{N}$ (Umegaki, 1954). The image C of this map is a unital subalgebra of $\mathcal{N}$ containing exactly those elements that commute with $C_k, k \in K$. Therefore, $E_c(A)C_k = E_c(AC_k) = C_k A C_k$ $(A \in \mathcal{N}, k \in K)$ for example. By means of this conditional expectation we can define the notion of the common cause system in the non-classical case:

Definition 5. A partition of the unit $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ is said to be the *common cause system* of the commuting events $A, B \in \mathcal{P}(\mathcal{N})$, which correlate in the state $\phi\colon \mathcal{N} \to \mathbf{C}$, if for those $k \in K$ for which $\phi(C_k) \neq 0$, the following condition holds:

$$\frac{(\phi \circ E_c)(ABC_k)}{\phi(C_k)} \quad = \quad \frac{(\phi \circ E_c)(AC_k)}{\phi(C_k)}\frac{(\phi \circ E_c)(BC_k)}{\phi(C_k)}. \tag{37}$$

If C_k commutes with both A and B for all $k \in K$, we call $\{C_k\}_{k\in K}$ a *commuting* common cause system, otherwise a *noncommuting* one. A common cause system of size $|K| = 2$ is called a *common cause*.

Some remarks are in place here. First, using the 'theorem of total probability' the common cause condition (37) can be written as

$$(\phi \circ E_c)(ABC_k))\,(\phi \circ E_c)(A^\perp B^\perp C_k) \quad = \quad (\phi \circ E_c)(AB^\perp C_k)\,(\phi \circ E_c)(A^\perp BC_k),\ k \in K. \quad (38)$$

One can even allow here the case $\phi(C_k) = 0$, since then both sides of (38) are zero.

Second, the non-classical character of the common cause system of Definition 5 lies in the fact that the common cause system need *not* commute with the correlating events. If the events A and B commute with $C_k, k \in K$, then not only $C_k \in \mathcal{C}$ but also $A, B, A^\perp, B^\perp \in \mathcal{C}$, and therefore $E_c(ABC_k) = ABC_k$, for example. Thus, the conditional expectation E_c vanishes from the defining equation (37); and (38) leads to

$$\phi(ABC_k)\,\phi(A^\perp B^\perp C_k) \quad = \quad \phi(AB^\perp C_k)\,\phi(A^\perp BC_k). \quad (39)$$

Finally, it is obvious from (39) that if $C_k \leq X$ with $X = A, A^\perp, B$ or $B^\perp$ for any $k \in K$ then $\{C_k\}_{k\in K}$ serve as a common cause system (and hence a commuting common cause system) of the given correlation independently of the chosen state ϕ. These solutions are called *trivial common cause systems*. In case of common cause, $|K| = 2$, triviality means that $\{C_k\} = \{A, A^\perp\}$ or $\{C_k\} = \{B, B^\perp\}$.

Having generalized the notion of the common cause system for the quantum case, the next step is to localize it. Suppose that the projection A is localized in the algebra $\mathcal{A}(V_A)$ with support V_A and the projection B is localized in the algebra $\mathcal{A}(V_B)$ with support V_B such that V_A'' and V_B'' are spacelike separated double cones in a spacetime $\mathcal{S}$. A common cause system $\{C_k\}_{k\in K}$ is said to be a *commuting/noncommuting (strong/weak) common cause system* of the correlation between A and B if $\{C_k\}_{k\in K}$ is localizable in an algebra $\mathcal{A}(V_C)$ with support V_C such that V_C is in $cpast(V_A, V_B)$ $(spast(V_A, V_B)/wpast(V_A, V_B))$.

In the same vein, we obtain the definition of the *joint common cause system* in the non-classical case. Let $\{(A_m, B_n); m \in M, n \in N\}$ be a set of pairs of commuting projections correlating in the sense that

$$\phi(A_m B_n) \quad \neq \quad \phi(A_m)\,\phi(B_n). \quad (40)$$

Definition 6. A partition of the unit $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ is said to be a *joint common cause system* of the set $\{(A_m, B_n); m \in M, n \in N\}$ of commuting pairs of correlating events, if for any $k \in K$, when $\phi(C_k) \neq 0$, the conditions

$$\frac{(\phi \circ E_c)(A_m B_n C_k)}{\phi(C_k)} \quad = \quad \frac{(\phi \circ E_c)(A_m C_k)}{\phi(C_k)}\,\frac{(\phi \circ E_c)(B_n C_k)}{\phi(C_k)}, \quad m \in M, n \in N \quad (41)$$

hold, where E_c is the conditional expectation defined in (36). Again, if $\{C_k\}_{k\in K}$ commutes with A_m and B_n for all $m \in M, n \in N$, then we call it a *commuting* joint common cause system, otherwise a *noncommuting* one.

Equation (41) can again be understood in the more permissive way as

$$(\phi \circ E_c)(A_m B_n C_k))\,(\phi \circ E_c)(A_m^\perp B_n^\perp C_k) \quad = \quad (\phi \circ E_c)(A_m B_n^\perp C_k)\,(\phi \circ E_c)(A_m^\perp B_n C_k) \quad (42)$$

incorporating cases when $\phi(C_k) = 0$.

And here comes a subtle point. Having introduced the notion of the joint common cause system of a correlation in the preceding Section we went over to conditional correlations and defined a *local,*

non-conspriratorial common causal explanation of these correlations. What is the analogue move in the non-classical case? We claim that we need *not* introduce any new concept; the definition of a *local, non-conspriratorial* common cause system in the non-classical case is just *identical* to the one given in Definition 6 that is to the definition of the joint common cause system. For the details see the Appendix (and (Butterfield 1995)). So from now on we drop the prefix 'local, non-conspiratorial' before the term 'joint common cause system' in the non-classical case.

Now, we are able to ask whether there is a proposition similary to Proposition 1 in the non-classical case, that is whether one can derive a CH inequality (2) from the fact that the set of correlating projections $\{(A_m, B_n); m \in M, n \in N\}$ has a *joint common causal explanation*? The following proposition provides a sufficient condition.

Proposition 2. Let $A_m \in \mathcal{A}(V_A)$ and $B_n \in \mathcal{A}(V_B)$ $(m, n = 1, 2)$ be four projections localized in spacelike separated spacetime regions V_A and V_B, respectively, which correlate in the locally faithful state ϕ in the sense of (40). Suppose that $\{(A_m, B_n); m, n = 1, 2\}$ has a joint common causal explanation in the sense of Definition 6. Then for any $m, m', n, n' = 1, 2; m \neq m'; n \neq n'$ the CH inequality

$$-1 \leqslant (\phi \circ E_c)(A_m B_n + A_m B_{n'} + A_{m'} B_n - A_{m'} B_{n'} - A_m - B_n) \leqslant 0. \tag{43}$$

holds for the state $\phi \circ E_c$. If the joint common cause is a *commuting* one, then the CH inequality holds for the original state ϕ:

$$-1 \leqslant \phi(A_m B_n + A_m B_{n'} + A_{m'} B_n - A_{m'} B_{n'} - A_m - B_n) \leqslant 0. \tag{44}$$

Proof. Substituting the expressions

$$\alpha := \frac{(\phi \circ E_c)(A_m C_k)}{\phi(C_k)} \tag{45}$$

$$\alpha' := \frac{(\phi \circ E_c)(A_{m'} C_k)}{\phi(C_k)} \tag{46}$$

$$\beta := \frac{(\phi \circ E_c)(B_n C_k)}{\phi(C_k)} \tag{47}$$

$$\beta' := \frac{(\phi \circ E_c)(B_{n'} C_k)}{\phi(C_k)} \tag{48}$$

into the inequality

$$-1 \leqslant \alpha\beta + \alpha\beta' + \alpha'\beta - \alpha'\beta' - \alpha - \beta \leqslant 0$$

and using (41) we get

$$\begin{aligned} -1 \leqslant & \frac{(\phi \circ E_c)(A_m B_n C_k)}{\phi(C_k)} + \frac{(\phi \circ E_c)(A_m B_{n'} C_k)}{\phi(C_k)} + \frac{(\phi \circ E_c)(A_{m'} B_n C_k)}{\phi(C_k)} \\ & - \frac{(\phi \circ E_c)(A_{m'} B_{n'} C_k)}{\phi(C_k)} - \frac{(\phi \circ E_c)(A_m C_k)}{\phi(C_k)} - \frac{(\phi \circ E_c)(B_n C_k)}{\phi(C_k)} \leqslant 0. \end{aligned} \tag{49}$$

Multiplying the above inequality by $\phi(C_k)$ and summing up for the index k one obtains

$$\begin{aligned} -1 \leqslant \sum_k \Big(& (\phi \circ E_c)(A_m B_n C_k) + (\phi \circ E_c)(A_m B_{n'} C_k) + (\phi \circ E_c)(A_{m'} B_n C_k) \\ & - (\phi \circ E_c)(A_{m'} B_{n'} C_k) - (\phi \circ E_c)(A_m C_k) - (\phi \circ E_c)(B_n C_k) \Big) \leqslant 0, \end{aligned} \tag{50}$$

which leads to (43) by performing the summation. If $\{C_k\}_{k\in K}$ is a *commuting* joint common cause system, then E_c drops out from the above expression since all the arguments are in $\mathcal{C}$ (see the remark before (38)). Therefore (50) becomes identical to (44), which completes the proof. ■

First note that similarly to Proposition 1, neither Proposition 2 refers to the spacetime localization of $\{C_k\}$ in a direct way. Indirectly, however, it restricts the localization of the possible joint common cause systems for states violating the CH inequality (44): the support of $\{C_k\}$ *must* intersect the union of the causal past or the causal future of $V_A \cup V_B$. It is so because otherwise the support of $\{C_k\}_{k\in K}$ would be spacelike separated from those of A and B, and hence $\{C_k\}$ would be a *commuting* joint common cause system for a set of correlations violating the CH inequality (44), in contradiction with Proposition 2.

Proposition 2—similarly to Proposition 1—provides a *common causal justification* of the CH inequality (44). It states that in order to yield a *commuting* joint common causal explanation for the set $\{(A_m, B_n); m, n = 1, 2\}$ the CH inequality (44) has to be satisfied. But what is the situation with *noncommuting* common cause systems? Since—apart from (43)—Proposition 2 is silent about the relation between a *noncommuting* joint common causal explanation and the CH inequality (44), the question arises: Can a *set* of correlations violating the CH inequality (44) have a *noncommuting* joint common causal explanation? Before addressing this question, we pose an easier one: Can *a single* correlation have a common causal explanation in AQFT? This leads us over to the question of the validity of the Common Cause Principles in AQFT.

5 Common Cause Principles in algebraic quantum field theory

Reichenbach's Common Cause Principle (CCP) is the following hypothesis: If there is a correlation between two events and there is no direct causal (or logical) connection between the correlating events, then there exists a common cause of the correlation. The precise definition of this informal statement that fits to the algebraic quantum field theoretical setting is the following:

Definition 7. A $\mathcal{P}_\mathcal{K}$-covariant local quantum theory $\{\mathcal{A}(V), V \in \mathcal{K}\}$ is said to satisfy the Commutative/Noncommutative (Weak/Strong) Common Cause Principle if for any pair $A \in \mathcal{A}(V_1)$ and $B \in \mathcal{A}(V_2)$ of projections supported in spacelike separated regions $V_1, V_2 \in \mathcal{K}$ and for every locally faithful state $\phi\colon \mathcal{A} \to \mathbf{C}$ establishing a correlation between A and B, there exists a *nontrivial* commuting/noncommuting common cause system $\{C_k\}_{k\in K} \subset \mathcal{A}(V), V \in \mathcal{K}$ of the correlation (34) such that the localization region V is in the (weak/strong) common past of V_1 and V_2.

What is the status of these six different notions of the Common Cause Principle in AQFT?

The question whether the Commutative Common Cause Principles are valid in a Poincaré covariant local quantum theory in the von Neumann algebraic setting was first raised by Rédei (1997, 1998). As an answer to this question, Rédei and Summers (2002, 2007) have shown that the Commutative Weak CCP is valid in algebraic quantum field theory with locally infinite degrees of freedom. Namely, in the von Neumann setting they proved that for every locally normal and faithful state and for every superluminally correlating pair of projections there exists a weak common cause, that is a common cause system of size 2 in the weak past of the correlating projections. They have also shown (Rédei and Summers, 2002, p 352) that the localization of a common cause $C < AB$ cannot be restricted to $wpast(V_1, V_2) \setminus I_-(V_1)$ or $wpast(V_1, V_2) \setminus I_-(V_2)$ due to logical independence of spacelike separated 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).[7] Hence, CCP is not valid there. Restricting ourselves to *local* algebras in Minkowski spaces the situation is not clear. We are of the opinion that one cannot decide on the validity of the (Strong) CCP without an explicit reference to the dynamics since there is no bounded region V in $cpast(V_1, V_2)$ (hence neither in $spast(V_1, V_2)$) for which isotony would ensure that $\mathcal{A}(V_1 \cup V_2) \subset \mathcal{A}(V'')$. But dynamics relates the local algebras since $\mathcal{A}(V_1 \cup V_2) \subset \mathcal{A}(V'' + t) = \alpha_t(\mathcal{A}(V''))$ can be fulfilled for certain $V \subseteq V'' \subset cpast(V_1, V_2)$ and for certain time translation by t.

Coming back to the proof of Rédei and Summers, the proof had a crucial premise, namely that the algebras in question are *von Neumann algebras of type III*. Although these algebras arise in a natural way in the context of Poincaré covariant theories, other local quantum theories apply von Neumann algebras of other type. For example, theories with locally finite degrees of freedom are based on finite dimensional (type I) local von Neumann algebras. This raised the question whether the Commutative Weak CCP is valid in other local quantum theories. To address the problem Hofer-Szabó and Vecsernyés (2012a) have chosen the local quantum Ising model (see Müller, Vecsernyés) having locally finite degrees of freedom. It turned out that the Commutative Weak CCP is *not valid* in the local quantum Ising model and it cannot be valid either in theories with locally finite degrees of freedom in general.

But why should we require commutativity between the common cause and its effects at all?

Commutativity has a well-defined role in any quantum theories: observables should commute to be simultaneously measurable. In AQFT commutativity of observables with spacelike separated supports is an axiom. To put it simply, commutativity can be required for events which can happen 'at the same time'. But cause and effect are typically *not* this sort of events. If one considers ordinary QM, one well sees that observables do not commute even with their own time translates in general. For example, the time translate $x(t) := U(t)^{-1} x U(t)$ of the position operator x of the harmonic oscillator in QM does *not* commute with $x \equiv x(0)$ for generic t, since in the ground state vector ψ_0 we have

$$\big[x, x(t)\big]\, \psi_0 = \frac{-i\hbar \sin(\hbar\omega t)}{m\omega} \psi_0 \not\equiv 0. \tag{51}$$

Thus, if an observable A is not a conserved quantity, that is $A(t) \neq A$, then the commutator $[A, A(t)] \neq 0$ in general. So why should the commutators $[A, C]$ and $[B, C]$ vanish for the events A, B and for their common cause C supported in their (weak/common/strong) past? We think that commuting common causes are only unnecessary reminiscense of their classical formulation. Due to their relative spacetime localization, that is due to the time delay between the correlating events and the common cause, it is also an unreasonable assumption.

Abandoning commutativity in the definition of the common cause is therefore a natural move. To our knowledge the first to contemplate the possibility of the noncommuting common causes were Clifton and Ruetsche (1999) in their paper criticizing Rédei (1997, 1998) who required commutativity from the common cause. They say: "[requiring commutativity] bars form candidacy to the post of common cause the vast majority of events in the common past of events problematically correlated" (p 165). And indeed, the benefit of allowing noncommuting common causes is that the noncommutative version of the result of Rédei and Summers can be regained: as it was shown in (Hofer-Szabó and Vecsernyés 2012b), by allowing common causes that do *not* commute with the correlating events, the Weak CCP can be proven in local UHF-type quantum theories.

Now, let us turn to our original question as to whether a *set* of correlations violating the CH inequality (2) can have a noncommuting *joint* common causal explanation in AQFT. Since our answer is provided in an AQFT with locally finite degrees of freedom, in the local quantum Ising model, we give a short and non-technical tutorial to this model in the next Section. (For more detail see (Hofer-Szabó, Vecsernyés, 2012c).)

[7] We thank David Malament for calling our attention to this point and the paper of Wald.

6 Noncommutative common causes for correlations violating the CH inequality

Consider a 'discretized' version of the two dimensional Minkowski spacetime $\mathcal{M}^2$ which is composed of minimal double cones $\mathcal{O}^m(t,i)$ of unit diameter with their center in (t,i) for $t,i \in \mathbb{Z}$ or $t,i \in \mathbb{Z}+1/2$. The set $\{\mathcal{O}^m_i, i \in \frac{1}{2}\mathbb{Z}\}$ of such minimal double cones with $t = 0, -1/2$ defines a 'thickened' Cauchy surface in this spacetime (see Fig. 3). The double cone $\mathcal{O}^m_{i,j}$ sticked to this Cauchy surface is defined to be the smallest double cone containing both $\mathcal{O}^m_i$ and $\mathcal{O}^m_j$: $\mathcal{O}^m_{i,j} := \mathcal{O}^m_i \vee \mathcal{O}^m_j$. Similarly, let $\mathcal{O}^m(t,i;s,j) := \mathcal{O}^m(t,i) \vee \mathcal{O}^m(s,j)$. The directed set of such double cones is denoted by $\mathcal{K}^m$, and the directed subset of it whose elements are sticked to a Cauchy surface is denoted by $\mathcal{K}^m_{CS}$. Obviously, $\mathcal{K}^m_{CS}$ will be left invariant by integer space translations and $\mathcal{K}^m$ will be left invariant by integer space and time translations.

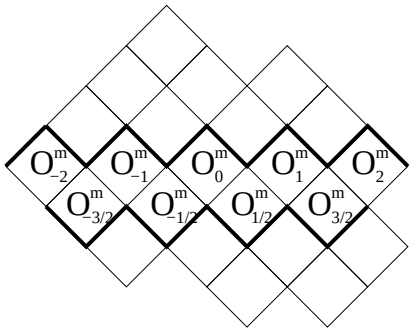


Figure 3: A thickened Cauchy surface in the two dimensional Minkowski space $\mathcal{M}^2$

The net of local algebras is defined as follows. The 'one-point' observable algebras associated to the minimal double cones $\mathcal{O}^m_i, i \in \frac{1}{2}\mathbb{Z}$ are defined to be $\mathcal{A}(\mathcal{O}^m_i) \simeq M_1(\mathbf{C}) \oplus M_1(\mathbf{C})$. Between the unitary selfadjoint generators $U_i \in \mathcal{A}(\mathcal{O}^m_i)$ one demands the following commutation relations:

$$U_i U_j = \begin{cases} -U_j U_i, & \text{if } |i-j| = \frac{1}{2}, \\ U_j U_i, & \text{otherwise.} \end{cases} \tag{52}$$

Now, the local algebras $\mathcal{A}(\mathcal{O}_{i,j}), \mathcal{O}_{i,j} \in \mathcal{K}^m_{CS}$ are linearly spanned by the monoms

$$U_i^{k_i} U_{i+\frac{1}{2}}^{k_{i+\frac{1}{2}}} \dots U_{j-\frac{1}{2}}^{k_{j-\frac{1}{2}}} U_j^{k_j} \tag{53}$$

where $k_i, k_{i+\frac{1}{2}} \dots k_{j-\frac{1}{2}}, k_j \in \{0,1\}$.[8]

Since the local algebras $\mathcal{A}(\mathcal{O}_{i,i-\frac{1}{2}+n})$, $i \in \frac{1}{2}\mathbb{Z}$ for $n \in \mathbb{N}$ are isomorphic to the full matrix algebra $M_{2^n}(\mathbb{C})$, the quasilocal observable algebra $\mathcal{A}$ is a uniformly hyperfinite (UHF) C^*-algebra and consequently there exists a unique (non-degenerate) normalized trace $\mathrm{Tr}\colon \mathcal{A} \to \mathbf{C}$ on it. We note that all nontrivial monoms in (53) have zero trace.

In order to extend the 'Cauchy surface net' $\{\mathcal{A}(\mathcal{O}), \mathcal{O} \in \mathcal{K}^m_{CS}\}$ to the net $\{\mathcal{A}(\mathcal{O}), \mathcal{O} \in \mathcal{K}^m\}$ in a causal and time translation covariant manner one has to classify causal (integer valued) time evolutions in the local quantum Ising model. This classification was given in (Müller, Vecsernyés) and it also was shown that the extended net satisfies isotony, Einstein causality, algebraic Haag

[8] For detailed Hopf algebraic description of the local quantum spin models see (Szlachányi, Vecsernyés, 1993), (Nill, Szlachányi, 1997), (Müller, Vecsernyés)).

duality

$$\mathcal{A}(\mathcal{O}')' \cap \mathcal{A} = \mathcal{A}(\mathcal{O}), \quad \mathcal{O} \in \mathcal{K}^m, \tag{54}$$

$\mathbf{Z} \times \mathbf{Z}$ covariance with respect to integer time and space translations and primitive causality:

$$\mathcal{A}(V) = \mathcal{A}(V''), \tag{55}$$

where V is a finite connected piece of a thickened Cauchy surface (composed of minimal double cones). V'' denotes the double spacelike complement of V, which is the smallest double cone in $\mathcal{K}^m$ containing V. We will be interested here only in a special subset of these causal automorphisms given by:

$$\beta(U_x) \quad = \quad U_{x-\frac{1}{2}} U_x U_{x+\frac{1}{2}}, \quad x \in \mathbf{Z} + \frac{1}{2}. \tag{56}$$

(In our following example we need not specify the choice for $\beta(U_x), x \in \mathbf{Z}$.) Now, consider the double cones $\mathcal{O}_A := \mathcal{O}^m(0,-1) \cup \mathcal{O}^m(\frac{1}{2},-\frac{1}{2})$ and $\mathcal{O}_B := \mathcal{O}^m(\frac{1}{2},\frac{1}{2}) \cup \mathcal{O}^m(0,1)$ and the 'two-point' algebras $\mathcal{A}(\mathcal{O}_A)$ and $\mathcal{A}(\mathcal{O}_B)$ pertaining to them. (See Fig. 4.) A linear basis of the algebra $\mathcal{A}(\mathcal{O}_A)$ is given

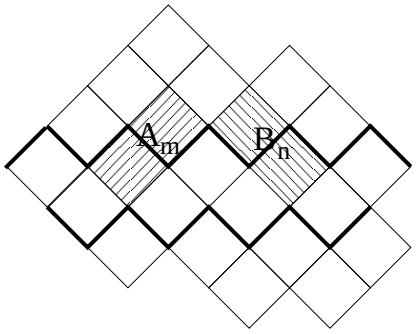


Figure 4: Projections in $\mathcal{A}(\mathcal{O}_A)$ and $\mathcal{A}(\mathcal{O}_B)$

by the monoms

$$\mathbf{1}, \qquad U_{-1}, \qquad \beta(U_{-\frac{1}{2}}) \equiv U_{-1} U_{-\frac{1}{2}} U_0, \qquad iU_{-1}\beta(U_{-\frac{1}{2}}) \equiv iU_{-\frac{1}{2}} U_0 \tag{57}$$

(where i in the fourth monom is the imaginary unit). They satisfy the same commutation relations like the Pauli matrices $\sigma_0 = \mathbf{1}, \sigma_x, \sigma_y$ and σ_z in $M_2(\mathbf{C})$. Therefore, introducing the notation

$$\mathbf{U} \quad := \quad (U_{-1},\, U_{-1}U_{-\frac{1}{2}}U_0,\, iU_{-\frac{1}{2}}U_0) \tag{58}$$

any minimal projection in $\mathcal{A}(\mathcal{O}_A)$ can be parametrized as

$$A(\mathbf{a}) \quad := \quad \frac{1}{2}\,(\mathbf{1} + \mathbf{aU}) \tag{59}$$

where $\mathbf{a} = (a_1, a_2, a_3)$ is a unit vector in $\mathbf{R}^3$. In the same vein, any minimal projection in $\mathcal{A}(\mathcal{O}_B)$ can be paremetrized as

$$B(\mathbf{b}) \quad := \quad \frac{1}{2}\,(\mathbf{1} + \mathbf{bV}) \tag{60}$$

where

$$\mathbf{V} \quad := \quad (U_1,\, -U_0U_{\frac{1}{2}}U_1,\, iU_0U_{\frac{1}{2}}) \tag{61}$$

is the vector composed of the generators of $\mathcal{A}(\mathcal{O}_B)$ and $\mathbf{b} = (b_1, b_2, b_3)$ is a unit vector in $\mathbf{R}^3$. The projections $A(\mathbf{a})$ and $B(\mathbf{b})$ can be interpreted as the event localized in $\mathcal{A}(\mathcal{O}_A)$ and $\mathcal{A}(\mathcal{O}_B)$, respectively pertaining to the generalized spin measurement in direction $\mathbf{a}$ and $\mathbf{b}$, respectively.

Now, consider two projections $A_m := A(\mathbf{a^m}); m = 1, 2$ localized in $\mathcal{O}_A$, and two other projections $B_n := B(\mathbf{b^n}); n = 1, 2$ localized in the spacelike separated double cone $\mathcal{O}_B$. Suppose that our system is in the faithful state $\phi(\,\cdot\,) = Tr(\rho\,\cdot\,)$ where

$$\rho \quad = \quad \rho(\lambda) := \mathbf{1} + \lambda\big(U_{-1}U_{-\frac{1}{2}}U_{\frac{1}{2}}U_1 - U_{-1}U_1 + U_{-\frac{1}{2}}U_{\frac{1}{2}}\big), \quad \lambda \in [0, 1). \tag{62}$$

For $\lambda = 1$ the state defined by (62) gives us back the usual singlet state. It is easy to see that in the state (62) the correlation between A_m and B_n will be:

$$corr(A_m, B_n) := \phi(A_mB_n) - \phi(A_m)\,\phi(B_n) = -\frac{\lambda}{4}\,\langle \mathbf{a^m}, \mathbf{b^n}\rangle \tag{63}$$

where $\langle\ ,\ \rangle$ is the scalar product in $\mathbf{R}^3$. In other words A_m and B_n will correlate whenever $\mathbf{a^m}$ and $\mathbf{b^n}$ are not orthogonal. Now, if $\mathbf{a^m}$ and $\mathbf{b^n}$ are chosen as

$$\mathbf{a^1} \quad = \quad (0, 1, 0) \tag{64}$$

$$\mathbf{a^2} \quad = \quad (1, 0, 0) \tag{65}$$

$$\mathbf{b^1} \quad = \quad \frac{1}{\sqrt{2}}(1, 1, 0) \tag{66}$$

$$\mathbf{b^2} \quad = \quad \frac{1}{\sqrt{2}}(-1, 1, 0) \tag{67}$$

the CH inequality (2) will be violated at the lower bound since

$$\phi(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1\big) =$$
$$-\frac{1}{2} - \frac{\lambda}{4}\left(\langle \mathbf{a^1}, \mathbf{b^1}\rangle + \langle \mathbf{a^1}, \mathbf{b^2}\rangle + \langle \mathbf{a^2}, \mathbf{b^1}\rangle - \langle \mathbf{a^2}, \mathbf{b^2}\rangle\right) = -\frac{1 + \lambda\sqrt{2}}{2}, \tag{68}$$

which is smaller than -1 if $\lambda > \frac{1}{\sqrt{2}}$. Or, equivalently, the CHSH inequality (1) where

$$X_m \quad := \quad 2A_m - \mathbf{1} \tag{69}$$

$$Y_n \quad := \quad 2B_n - \mathbf{1} \tag{70}$$

will be violated for the above setting since

$$\phi(X_1(Y_1 + Y_2) + X_1(Y_1 - Y_2)) =$$
$$= -\lambda\left(\langle \mathbf{a^1}, \mathbf{b^1} + \mathbf{b^2}\rangle + \langle \mathbf{a^2}, \mathbf{b^1} - \mathbf{b^2}\rangle\right) = -\lambda 2\sqrt{2} \tag{71}$$

is smaller than -2 if $\lambda > \frac{1}{\sqrt{2}}$. Both the CH and the CHSH inequality are maximally violated for the singlet state, that is if $\lambda = 1$.

The question whether the four correlations $\{(A_m, B_n); m, n = 1, 2\}$ violating the CH inequality (2) have a joint common causal explanation was answered in (Hofer-Szabó, Vecsernyés, 2012c) by the following

Proposition 3. Let $A_m := A(\mathbf{a^m}) \in \mathcal{A}(\mathcal{O}_A), B_n := B(\mathbf{b^n}) \in \mathcal{A}(\mathcal{O}_B); m,n = 1,2$ be four projections defined in (59)-(60), where $\mathbf{a^m}$ and $\mathbf{b^n}$ are non-orthogonal unit vectors in $\mathbf{R}^3$ establishing four correlations $\{(A_m, B_n); m,n = 1,2\}$ in the state (62). Let furthermore C be any projection localized in $\mathcal{O}_C := \mathcal{O}_{-\frac{1}{2}} \vee \mathcal{O}_{\frac{1}{2}} \in \mathcal{K}^m_{CS}$ (see Fig. 5.) of the shape

$$\begin{aligned} C &= \frac{1}{4}\left(\mathbf{1} + U_{-\frac{1}{2}}U_{\frac{1}{2}}\right)\left(\mathbf{1} + c_1U_0 + c_2U_{\frac{1}{2}} + c_3iU_0U_{\frac{1}{2}}\right) \\ &\quad + \frac{1}{4}\left(\mathbf{1} - U_{-\frac{1}{2}}U_{\frac{1}{2}}\right)\left(\mathbf{1} + c'_1U_0 + c'_2U_{\frac{1}{2}} + c'_3iU_0U_{\frac{1}{2}}\right) \end{aligned} \tag{72}$$

where $\mathbf{c} = (c_1, c_2, c_3)$ and $\mathbf{c'} = (c'_1, c'_2, c'_3)$ are arbitrary unit vectors in $\mathbf{R}^3$. Then $\{C, C^\perp\}$ is a joint common cause of the correlations $\{(A_m, B_n)\}$ if $a^m_3 b^n_3 = 0$ for any $m,n = 1,2$ and $c_2 = 0$.

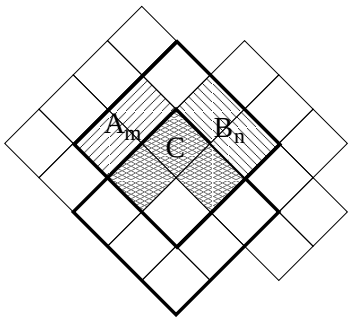


Figure 5: Localization of a common cause for the correlations $\{(A_m, B_n)\}$.

Since for the directions $\mathbf{a^m}$ and $\mathbf{b^n}$ defined in (64)-(67) the requirement $a^m_3 b^n_3 = 0$ holds for any $m,n = 1,2$, therefore the correlations (maximally) violating the CH/CHSH inequality *do* have a joint common cause—any C of form (72) with $c_2 = 0$.

Finally, here is a Proposition (consistently with the derivability of a CH inequality from the commuting joint common cause system) claiming that there exists no *commuting* joint common cause for these correlations even without any restriction to their localization (Hofer-Szabó, Vecsernyés, 2012c):

Proposition 4. Let $A_m \in \mathcal{A}(\mathcal{O}_A), B_n \in \mathcal{A}(\mathcal{O}_B); m,n = 1,2$ be projections defined in (59)-(60) with $\mathbf{a^m}$ and $\mathbf{b^n}$ given in (64)-(67). The correlations $\{(A_m, B_n); m,n = 1,2\}$ in the state (62) do *not* have a *commuting* joint common cause $\{C_1, C_2\}$ in $\mathcal{A}$.

Proposition 3 answers the question raised at the end of the last Section as to whether there is a common causal justification of the CH inequalities in the general, that is in the noncommuting case. The answer to this question is clearly *no*. The violation of the CH inequality for a given set of correlation does *not* prevent us from finding a common causal explanation for them. All we have to do is to extend our scope of search and to embrace noncommuting common causes in the common causal explanation. So the Bell inequalities in the non-classical case do *not* play the same role as in the classical one. In the classical case there was a direct logical link between the possibility of a common causal explanation and the validity of the Bell inequalities; here the violation of the Bell inequalities excludes only a subset of the possible common causal explanations containing the commuting ones. To put it differently, taking seriously the ontology of AQFT where events are represented by not necessarily commuting projections, one can provide a common causal explanation in a much wider range than simply sticking to commutative common causes.

7 On the meaning of noncommuting common causes

But what are the consequences of applying noncommutative common causes? Let us see the story from the beginning, going back to Reichenbach's original definition of the common cause. The Reichenbachian common cause has the nice property that the presence of a common cause implies a (positive) correlation between the events in question. This fact is a simple consequence of the following identity:

$$p(A \wedge B) - p(A)\,p(B) = p(C)p(C^{\perp})\big[p(A|C) - p(A|C^{\perp})\big]\big[p(B|C) - p(B|C^{\perp})\big]. \tag{73}$$

It is straightforward to check that if C is a Reichenbachian common cause fulfilling requirements (6)-(9) then the right hand side of (73) is positive therefore there is a positive correlation between A and B. In this sense the common cause provids a Hempelian explanation for the correlation.[9] Going over to the notion of the common cause system this 'explanatory force' of the common cause disappears: from the presence of the common cause (11) the correlation (10) between A and B does *not* follow. (For an attempt to define the notion of the common cause system such that it preserves this deductive relation between the common cause system and the correlation see (Hofer-Szabó and Rédei 2004, 2006).)

The noncommutative generalization of the common cause system is one step further into the direction of relaxing the relation between the common cause and the correlation. Here not only the deductive relation between the common cause and the correlation gets lost, but also the relation between the conditioned and unconditioned probalitity of the correlating events. Namely,

$$\phi(A) = \phi_c(A) := (\phi \circ E_c)(A) \equiv \sum_k \frac{(\phi \circ E_c)(AC_k)}{\phi(C_k)}\,\phi(C_k) \tag{74}$$

holds in general iff $A = E_c(A)$, that is iff $[A, C_k] = 0$ for all $k \in K$. That is the state ϕ_c differs from ϕ for $A \in \mathcal{A} \setminus \operatorname{Im} E_c$ in general, which means that the statistics of A can differ depending on whether we calcutate it directly from the state ϕ or as a weighted average of conditional probabilities over the subensembles C_k.

But then one might come up with the following concern: Noncommuting common causes are *not actual* but only *contrafactual* entities since if the C_k-s *had been realized*, then we *would have ended up* with another probability (the right hand side of (74)) for the correlating events than the actual ones (the left hand side of (74)). So these common causes cannot be realized in the same (actual) world in which those event are accomodated which they are supposed to explain.

We do not consider this objection to be serious against the application of noncommuting common causes. An analogy between the notion of the *common cause* and the notion of the *cause* in QM might help to illuminate why. An observable/event X can be said to be the *cause* of another observable/event Y in QM, if X evolves in time into Y. But if X and Y do not commute, then had X been earlier realized, the unitary dynamics would have been distorted, so X would not have evolved into Y. Still, we regard X to be the cause of Y. Similarly, C is a *common cause* of A and B if conditioned on it the correlation between A and B disappears. If C does not commute with A and B, then had C been realized, the statistics would have been distorted, so the probability of A, B and AB would be different. Still, we think that C is the common cause.

What is important to see here is that the definition of the common cause does *not* contain the requirement (which our classicaly informed intuition would dictate) that the conditional probabilites, when added up, should give back the unconditional probabilities, that is $\phi = \phi_c$ should fulfil. Or in other words, that the probability of the correlating events should be built up from a finer description of the situation provided by the common cause. To put it in a more formal way: the theorem of

[9] One is tempted to speculate that this desired property might just have been the reason why Reichenbach took up the statistical relevancy conditions (8)-(9) in the definition of the common cause.

total probability is *not* part of the definition of the common cause.[10] The defining property of the common cause is simply the *screening-off*.

So common causes might not be realized without the distortion of the statistics of the original correlating events. But this fact is ubiquitous for noncommuting observables in QM. If we tolerate this fact in general, then why not to tolerate it for common causes? As we have seen, allowing noncommuting common causes helps us to maintain Bell's original intuition concerning local causality.

8 Conclusions

In the paper we saw that the Bell inequalities used in AQFT cannot be given a common causal justification similar to the classical Bell inequalities if we allow noncommuting common causes in the explanation. Just the opposite is true: for a set of correlations violating the CH inequalities a noncommutative common causal explanation can be given and this common cause can be localized in the common past of the correlating events. Thus, abandoning commutativity gives us extra freedom in the search of common causes for correlations. But how big is this freedom? Is it big enough to find a common cause for *any set* of correlations? We saw that for the worst candidate, so to say, for the set *maximally* violating the CH inequality we have found such a common cause. But does it mean that this strategy can be applied across the board? What is the range of correlations for which a joint common causal explanation can be given? Is this range determined only by the size of the set of correlations or by some other properties thereof? Is it true for example that for any finite set of correlations a weak joint common causal explanation can always be given? Or to put it in a more formal way, can one always find a partition of the unit for any finite set of correlations such that the necessary condition (43) for a joint common causal explanation fulfills? All these questions are still open.

Appendix: In what sense non-classical joint common cause systems are local and non-conspiratorial?

In Section 4 we claimed that Definition 6 of the joint common cause system is the correct non-classical generalization of Definition 4 of the (classical) *local, non-conspiratorial* joint common cause system. But how can the single non-classical screening-off condition (41) generalize not only the classical screening-off condition (15) but also the locality conditions (16)-(17) and non-conspiracy (18)? This is the question we address in this Appendix.

Let us first introduce a classical probability measure p_{C_k} on a common measure space (Ω, Σ) for every element of a classical common cause system $\{C_k, k \in K\}$, if $p(C_k) \neq 0$:

$$p_{C_k}(X|x) \quad := \quad \frac{p(X \wedge C_k|x)}{p(C_k)}. \tag{75}$$

With this denotation screening-off (15), locality (16)-(17), and no-conspiracy (18) will read as

$$p_{C_k}(A_m \wedge B_n|a_m \wedge b_n) = p_{C_k}(A_m|a_m \wedge b_n)\, p_{C_k}(B_n|a_m \wedge b_n), \tag{76}$$

$$p_{C_k}(A_m|a_m \wedge b_n) = p_{C_k}(A_m|a_m \wedge b_{n'}), \tag{77}$$

$$p_{C_k}(B_n|a_m \wedge b_n) = p_{C_k}(B_n|a_{m'} \wedge b_n), \tag{78}$$

$$p_{C_k}(\Omega|a_m \wedge b_n) = 1, \tag{79}$$

[10] As it is not part of the definition of the *cause* either: if one measures X, one cannot reconstruct the probability of a noncommuting Y from the conditional probabities over the subensembles pertaining to the outcomes of X.

if one uses no-conspiracy (18) in the first three equations. The subscript C_k of the probability measure might remind the reader to the standard hidden variable approach where a parameter λ is used to index a set of probability measures on a common event algebra. In this approach the derivation of the Bell inequalities then proceeds through the summation/integration over this parameter. In our opinion this indexical treatment of the common cause conceals an important fact, namely that the common cause and the correlating events stand on the same ontological footing: they are all *events*, accomodated in a common event algebra with a single probability measure. Therefore the index in (76)-(79) is simply an abbreviation of the conditionalization (75), which abbreviation is motivated by trying to find a classically equivalent form, where the non-classicaly meaningless expression $a_m \wedge b_n \wedge C_k$ of non-commuting quantities can have a definite interpretation. (See below.)

Now, how does the non-classical Definition 6 of the joint common cause system relate to the above characterization of a classical *local, non-conspiratorial* joint common cause system? The link is provided by the (in our oppinion) correct interpretation of the non-classical probabilities according to which quantum probabilities are *classical conditional probabilities*. The quantum probability $\phi(X)$ of a projection X is to be interpreted as a *conditional* probability $p(X_{cl}|x_{cl})$ of getting the outcome X_{cl} *given* the quantity x_{cl} has been set to be measured. The precise mathematical formulation of this interpretation is given in the so-called 'Kolmogorovian Censorship Hyptothesis'. Here we just state the proposition; for the proof see (Bana and Durt 1997), (Szabó 2001) and (Rédei 2010).

Kolmogorovian Censorship Hypothesis. Let $(\mathcal{N}, \mathcal{P}(\mathcal{N}), \phi)$ be a non-classical probability space. Let Γ be a countable set of non-commuting selfadjoint operators in $\mathcal{N}$. For every $Q \in \Gamma$, let $\mathcal{P}(Q)$ be a maximal Abelian sublattice of $\mathcal{P}(\mathcal{N})$ containing all the spectral projections of Q. Finally, let a map $p_0 : \Gamma \to [0,1]$ be such that

$$\sum_{Q \in \Gamma} p_0(Q) = 1, \qquad p_0(Q) > 0. \tag{80}$$

Then there exists a classical probability space (Ω, Σ, p) such that for every projection X^Q in any $\mathcal{P}(Q)$ there exist events X^Q_{cl} and x^Q_{cl} in Σ such that

$$X^Q_{cl} \subset x^Q_{cl} \tag{81}$$

$$x^Q_{cl} \cap x^R_{cl} = 0, \quad \text{if } Q \neq R \tag{82}$$

$$p(x^Q_{cl}) = p_0(Q) \tag{83}$$

$$\phi(X^Q) = p(X^Q_{cl}|x^Q_{cl}) \tag{84}$$

The intuitive content of the above proposition is the following. A set of incompatible observables represented by noncommuting selfadjoint operators in the set Γ are selected for measurement with the probabilities $p_0(Q)$ specified in (80). This measurement and selection procedure is then represented by classical events X^Q_{cl} and x^Q_{cl}, respectively: X^Q_{cl} represents a certain measurement outcome of the measurement Q, and x^Q_{cl} is the classical event of setting up the measurement device to measure Q. Condition (81) expresses that no outcome is possible without this setting up of a measuring device. Condition (82) expresses that incompatible observables Q and R cannot be simultaneously measured: the measurement choices x^Q_{cl} and x^R_{cl} are disjoint events. Condition (83) states that the classical probability model captures the prescribed probabilities $p_0(Q)$ as the probability of the measurement choices. Finally, condition (84) is the central relation of the Hypothesis, it states that quantum probabilities can be written as classical conditional probabilities: conditional probabilities of outcomes of measurements on condition that the appropriate measuring device has been set up.

Applying the above proposition to our case,[11] we obtain that the quantum probabilities $\phi(A_m)$,

[11] From now on, we will denote both the classical event and the projection representing it by the same symbol. However, the quantum state ϕ or the classical probability p will always indicate in which sense we use it.

$\phi(B_n)$ and $\phi(A_mB_n)$ can be interpreted as classical conditional probabilities $p(A_m|a_m)$, $p(B_n|b_n)$ and $p(A_m \wedge B_n|a_m \wedge b_n)$, respectively, with A_m, B_n, a_m and b_n $(m \in M, n \in N)$ accomodated in a classical probability space (Ω, Σ, p). Hence the quantum correlations

$$\phi(A_mB_n) \quad \neq \quad \phi(A_m)\,\phi(B_n) \tag{85}$$

between the elements of the set $\{(A_m, B_n); m \in M, n \in N\}$ can be interpreted as *conditional* correlations

$$p(A_m \wedge B_n \,|\, a_m \wedge b_n) \quad \neq \quad p(A_m|a_m)\,p(B_n|b_n) \tag{86}$$

between classical measurement outcome events conditioned on measurement choice events in accordance with (14).

To see the link between the classical and non-classical version of the common cause let us first introduce a similar notation for the conditionalization on C_k in the non-classical case, if $\phi(C_k) \neq 0$, as was introduced above in (75) for the classical case, that is let

$$\phi_{C_k}(X) \quad := \quad \frac{(\phi \circ E_c)(XC_k)}{\phi(C_k)} = \frac{\phi(C_kXC_k)}{\phi(C_k)}. \tag{87}$$

With this notation the definition of the non-classical joint common cause system reads as follows:

$$\phi_{C_k}(A_mB_n) \quad = \quad \phi_{C_k}(A_m)\,\phi_{C_k}(B_n). \tag{88}$$

Using the Kolmogorovian Censorship Hypothesis the classical interpretation of (88) is the following:

$$p_{C_k}(A_m \wedge B_n|a_m \wedge b_n) \quad = \quad p_{C_k}(A_m|a_m)\,p_{C_k}(B_n|b_n) \tag{89}$$

which is *almost* the screening-off (76) except that the conditions on the right hand side are *not* $a_m \wedge b_n$. This defect will be cured however by the locality conditions. Observe namely that since A_m and B_n commute, therefore

$$\phi_{C_k}(A_m) \quad = \quad \phi_{C_k}(A_mB_n) + \phi_{C_k}(A_mB_n^{\perp}) \tag{90}$$

$$\phi_{C_k}(B_n) \quad = \quad \phi_{C_k}(A_mB_n) + \phi_{C_k}(A_m^{\perp}B_n) \tag{91}$$

which translated into classical conditional probabilities due to the Kolmogorovian Censorship Hypothesis read as:

$$p_{C_k}(A_m|a_m) \quad = \quad p_{C_k}(A_m \wedge B_n|a_m \wedge b_n) + p_{C_k}(A_m \wedge B_n^{\perp}|a_m \wedge b_n) = p_{C_k}(A_m|a_m \wedge b_n) \tag{92}$$

$$p_{C_k}(B_n|b_n) \quad = \quad p_{C_k}(A_m \wedge B_n|a_m \wedge b_n) + p_{C_k}(A_m^{\perp} \wedge B_n|a_m \wedge b_n) = p_{C_k}(B_n|a_m \wedge b_n) \tag{93}$$

Now, observe that (92)-(93) are equivalent to locality (77)-(78), so locality is 'automatically' fulfilled for the non-classical common cause due to the commutativity of A_m and B_n. (This fact is sometimes referred as the 'no-signalling theorem'; for more on that see (Schlieder 1969).) Moreover (92)-(93) also cure the defect of (89), since

$$p_{C_k}(A_m|a_m)\,p_{C_k}(B_n|b_n)$$

on the right hand side of (89) can be replaced with

$$p_{C_k}(A_m|a_m \wedge b_n)\,p_{C_k}(B_n|a_m \wedge b_n)$$

turning (89) into the classical screening-off property (76).

Putting all this together, a *non-classica*l, local, non-conspiratorial joint common causal explanation of the correlations (85) is a partition $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ if for any $k \in K$ the following requirements hold:

$$\phi_{C_k}(A_m B_n) = \phi_{C_k}(A_m)\,\phi_{C_k}(B_n) \quad (94)$$

$$\phi_{C_k}(A_m) = \phi_{C_k}(A_m B_n) + \phi_{C_k}(A_m B_n^{\perp}) \quad (95)$$

$$\phi_{C_k}(B_n) = \phi_{C_k}(A_m B_n) + \phi_{C_k}(A_m^{\perp} B_n) \quad (96)$$

$$\phi_{C_k}(\mathbf{1}) = 1. \quad (97)$$

which using the Kolmogorovian Censorship Hypothesis as a 'translation manual' leads us over to the *classical*, local, non-conspiratorial joint common causal explanation (76)-(79) of the correlations (86). But recall that (95)-(97) representing locality and no-conspiracy are just *identities*, and hence the screening-off condition (94) carries the whole content of the common causal explanation—in accordance with our Definition 6.

Acknowledgements. We wish to thank Jeffrey Bub, John D. Norton, David Malament, Jim Woodward, the Southern California Philosophy of Physics Group and the Foundations of Physics Discussion Group at the University of Maryland for helpful comments on an earlier version of the paper. This work has been supported by the Hungarian Scientific Research Fund, OTKA K-68195 and OTKA K-100715 and by the Fulbright Research Grant while G. H-Sz. was a Visiting Fellow at the Center for Philosophy of Science in the University of Pittsburgh.

References

G. Bacciagaluppi, "Separation theorems and Bell inequalities in algebraic QM," *Symposium on the Foundations of Modern Physics 1993: Quantum Measurement, Irreversibility and Physics of Information*, World Scientific, 29-37 (1994).

G. Bana and T. Durt, "Proof of Kolmogorovian Censorship," *Foundations of Physics*, **27**, 1355–1373. (1997).

J. S. Bell, "Beables for quantum field theory," (TH-2053-CERN, presented at the Sixth GIFT Seminar, Jaca, 2–7 June 1975) reprinted in J. S. Bell, *Speakable and Unspeakable in Quantum Mechanics*, (Cambridge: Cambridge University Press, 1987, 52-62.).

I. Bengtson and K. Życzkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge University Press, Cambridge, 2006.

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.

D. Buchholz, C. D'Antoni and K. Fredenhagen, "The universal structure of local algebras," *Commun. Math. Phys.*, **111/1**, 123-135 (1987).

J. F. Clauser, M.A. Horne, A. Shimony and R. A. Holt, "Proposed experiment to test local hidden-variable theories," *Phys. Rev. Lett.*, **23**, 880-884 (1969).

J. F. Clauser and M. A. Horne, "Experimental consequences of objective local theories," *Phys. Rev. D*, **10**, 526-535 (1974).

R. Clifton and L. Ruetsche, "Changing the subject: Redei on causal dependence and screening off in relativistic quantum field theory," *Philosophy of Science*, **66**, S156-S169 (1999).

K. Fredenhagen, "On the modular structure of local algebras of observables" *Commun. Math. Phys.*, **97**, 79-89 (1985).

N. Gisin, *Quantum Reality, Relativistic Causality, and Closing the Epistemic Circle*, (The Western Ontario Series in Philosophy of Science, **73**, III / 1, 125-138, 2009).

C. Glymour, "Markov properties and quantum experiments," in W. Demopoulos and I. Pitowsky (eds.) *Physical Theory and its Interpretation*, (Springer, 117-126, 2006).

R. Haag, *Local Quantum Physics*, (Springer Verlag, Berlin, 1992).

H. Halvorson and R. Clifton, "Generic Bell correlation between arbitrary local algebras in quantum field theory," *J. Math. Phys.*, **41**, 1711-1717 (2000).

H. Halvorson, "Algebraic quantum field theory," in J. Butterfield, J. Earman (eds.), *Philosophy of Physics, Vol. I*, Elsevier, Amsterdam, 731-922 (2007).

J. Henson (2005). "Comparing causality principles," *Studies in the History and Philosophy of Modern Physics*, **36**, 519–543.

G. Hofer-Szabó, G., M. Rédei (2004). "Reichenbachian Common Cause Systems," *International Journal of Theoretical Physics*, **34**, 1819-1826.

G. Hofer-Szabó, G., M. Rédei (2006). "Reichenbachian Common Cause Systems of Arbitrary Finite Size Exist," *Foundations of Physics Letters*, **35**, 745-746.

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, P. Vecsernyés, "Noncommutative Common Cause Principles in AQFT," *Found. Phys.*, (submitted) (2012b).

G. Hofer-Szabó, P. Vecsernyés, "Noncommuting local common causes for correlations violating the Clauser–Horne inequality," *Journal of Physics A*, (submitted) (2012c).

J. Jarrett, "On the Physical Significance of the Locality Conditions in Bell Arguments," *Nous*, **18**, 569-589, (1984).

L. Landau, "On the violation of Bell's inequality in quantum theory," *Phys. Lett. A* **120**, 54-56 (1987).

G. Lüders, "Über die Zustandsänderung durch den Messprozess," *Annalen der Physik* **8**, 322-328 (1951)..

V.F. Müller and P. Vecsernyés, "The phase structure of G-spin models", *to be published*

F. Nill and K. Szlachányi, "Quantum chains of Hopf algebras with quantum double cosymmetry" *Commun. Math. Phys.*, **187** 159-200 (1997).

M. Rédei, "Reichenbach's Common Cause Principle and quantum field theory," *Found. Phys.*, **27**, 1309–1321 (1997).

M. Rédei, *Quantum Logic in Algebraic Approach*, (Kluwer Academic Publishers, Dordrecht, 1998).

M. Rédei, "Kolmogorovian Censorship Hypothesis for general quantum probability theories," *Manuscrito - Revista Internacional de Filosofia*, **33**, 365–380 (2010).

M. Rédei and J. S. Summers, "Local primitive causality and the Common Cause Principle in quantum field theory," *Found. Phys.*, **32**, 335-355 (2002).

M. Rédei and J. S. Summers, "Remarks on Causality in relativistic quantum field theory," *Int. J. Theor. Phys.*, **46**, 2053–2062 (2007).

H. Reichenbach, *The Direction of Time*, (University of California Press, Los Angeles, 1956).

A. Shimony, "Events and processes in the quantum world," in Penrose, R. and Isham, C. (eds.), *Quantum concepts in space and time*, 182–203. (University Press, Oxford, 1986.)

S. Schlieder, "Einige Bemerkungen über Projektionsoperatoren (Konsequenzen eines Theorems von Borchers," *Communications in Mathematical Physics*, **13**, 216-225 (1969).

S. J. Summers, "On the independence of local algebras in quantum field theory" *Reviews in Mathematical Physics*, **2**, 201-247 (1990).

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).

L. E. Szabó, "Quantum structures do not exist in reality," *Int. J. of Theor. Phys*, **37**, 449-456. (1998)

L. E. Szabó, "Critical reflections on quantum probability theory," in Rédei and Stölzner (eds.) *John von Neumann and the Foundations of Quantum Physics.* (Institute Vienna Circle Yearbook. Dordrecht: Kluwer Academic Publishers, 201–219, 2001)

K. Szlachányi and P. Vecsernyés, "Quantum symmetry and braid group statistics in G-spin models" *Commun. Math. Phys.*, **156**, 127-168 (1993).

H. Umegaki, "Conditional expectation in an operator algebra, I," *Tohoku Math. J. (2)*, **6/2-3**, 177-181 (1954).

B. C. Van Fraassen, "Rational belief and common cause principle," in: R. McLaughlin (ed.), *What? Where? When? Why?*, Reidel, 193-209 (1982).

R. M. Wald, "Correlations beyond the horizon," *General Relativity and Gravitation*, **24**, 1111-1116 (1992).

For *Pacific APA*

Date: Monday, March 19, 2012
Target Word Count: approximately 6000 words
Actual Word Count: 6941

What kinds of kind are the senses?

— Brian L. Keeley —

Philosophy and *Science, Technology & Society* Field Groups,
Pitzer College

ABSTRACT

In Western common sense, one speaks of there being five human senses, a claim apparently challenged by the biological and psychological sciences. Part of this challenge comes in the form of claiming the existence of additional senses (proprioception, pain, a human pheromone sense). Part of the challenge comes from positing multiple senses where common sense only speaks of one, such as with the fractionation of "touch" into pressure and temperature senses. One conceptual difficulty in thinking about the number and division of senses is that it's not clear whether the different senses constitute natural kinds and, if not, what kind of kind they are. Should we favor antirealism with respect to the senses, akin to the arguments of some concerning the nature of species or race? I will argue that this first problem is compounded by another: that we ought to be pluralists with respect to the senses—what is meant by the term "sense" varies from context to context, varying even between scientific contexts.

I. Introduction

In a recent paper, "The senses as psychological kinds," Matthew Nudds (2011) observes and asks, "We see, hear, touch, smell, and taste things. In distinguishing determinate ways of perceiving things, what are we distinguishing between? What, in other words, is a sense modality" (311)? He goes on to note that there are many differences to be found between the senses, but asks, "...which, if any, of these differences are those that *really matter*?" (311, my emphasis). This

is all just a way of asking a question about the metaphysical nature of the senses.

At first glance, it might seem that the difference between the senses would be a paradigm case of a difference in *kind*. On many different commonsense criteria, seeing and hearing, say, are fundamentally different: They are carried out by different organs (eyes and ears, respectively). They bring us information about very different aspects of the world (colors and pitches). They involve the transduction of different kinds of energy (electromagnetic and mechanical). Further, the experiences of one are not easily confused with those of the other. The differences here would seem to be *brute*; a starting point for further analysis, not something open for much analysis itself. Apparently, seeing and hearing are *just different.*

In exploring these questions, Nudds considers the possibility that the senses are natural kinds—more precisely: "psychological kinds"—and are differentiated in virtue of the different perceptual (psychological) mechanisms that operate in the cases of the differing senses. However, although he sees such an account as having merit, he is not ultimately willing to bet on it.[1] Instead, he proposes that,

> ...we could accept that in distinguishing different perceptions we are distinguishing them on the basis of how they were produced but give up on the idea that we can explain or give an account of the different ways that perceptions are produced that is independent of our practice of making the distinction.

[1] He's bothered that embracing such an account very likely will lead to the conclusion that the claim of common sense that humans have five (and exactly five) senses will be refuted. He finds this anathema and would prefer to avoid embracing such an eliminative materialist line of reasoning. As they say, one person's *modus ponens* is another person's *modus tolens.*

> According to this approach, all visual perceptions are produced in the same way, and different ways of perceiving are individuated relative to a social practice of explaining and understanding behavior. On this view a sense modality is what might be called a *social* kind rather than a natural kind. Such an account may provide the best account of what a sense modality, as we commonly understand it, actually is. (338, emphasis in original)

Nudds is exploring an interesting issue: are the senses natural kinds? Social kinds? My own reaction is that there are deeper issues here of which he is only scratching the surface. Nudds' paper is one of the only ones I am familiar with that grapples with the question of what it is that we are presupposing in the first place when we talk of different senses.[2] But even Nudds' analysis is pretty slim on what the options are. For example, the paragraph I quote above (the final paragraph of the paper) is pretty much all he says on what a "social kind" is supposed to be. Similarly, although the notion of "natural kinds" is central to the paper, Nudds has almost nothing to say about what he takes that concept to mean; it is left largely unpacked. He draws no connection to the literature in philosophy of science over the metaphysical nature of natural vs. other kinds.

Further, we shouldn't take it for granted that the senses are, in fact, properly thought of as different kinds, whether paradigmatically or not. This oversight is important because, as I will show in this paper, it is far from clear what kind of kind the senses are, if they are any "kind" at all. Further, I will argue that what kind the senses are, in fact, varies from context to context. Treating "the senses" uniformly as kinds of a particular kind confuses rather than clarifies the situation.

---

[2] Nudds' previous paper in 2004 is another.

II. Why does the question matter?

Why does the question concerning the kind-hood of senses matter? So what? First, the questions here are centrally important to the metaphysics of mind. Ought we be realists with respect to the senses? Is there some fact of the matter about, say, how many senses species-typical humans have and what they are? If not, what should we say about the metaphysical standing of the individual senses? To get the idea of what might be at stake here, let me point to some related metaphysical questions.

Species: Species are clearly a central ontological category within the science of biology—it is no accident that Darwin's book was entitled *On the Origin of Species*. It is important, therefore, to understand the metaphysical nature of this central concept, and this in turn is a vexed issue in both philosophy of biology as well as biology itself. Recently, much ink has been spilled over whether we ought to be realists with respect to species, as well as the closely related question of what is the nature of the species concept.[3] Darwin can be understood as overthrowing the *essentialist* understanding of species that had reigned in biology at least since Aristotle. Given that species evolve, as Darwin showed us, then they cannot have the immutable, God-given essences non-evolutionary models proposed. These days, there are those, such as Hull (1978) and Ghiselin (1974), who argue that species are best thought of as spatiotemporally extended *individuals*. Others, such as Kitcher (1984; 1989), argue that species are *sets*, while others (e.g., Boyd (1999), Griffiths (1999)) argue that they are *homeostatic*

[3] See Wilson (1999) for a collection of essays on these topics.

property clusters. Still others are one form or another of antirealists about species, e.g., (Stanford 1995) and arguably even Darwin himself. All of this discussion is made relevant, in part, because of the central role that the concept *species* plays in the field. It seems only right that we know exactly what sort of thing we're talking about when biologists develop their theories.[4]

Race: Where *species* is centrally important to biology, race is clearly an important social category, for better or for worse. For many of the same reasons as with species, it is crucial to understand the nature of the category of race. To what extent are racial categories "biologically real"? To what extent are they "socially constructed"? Again, this is a topic of ongoing controversy and discussion. As with species, for centuries race was conceived in essentialist terms, a view K. Anthony Appiah (1996) calls "racialism". Appiah rejects the biological reality of races, arguing instead that race categories are best thought of as identities that individuals chose to take on or which are culturally imposed on them. Others, such as Kitcher (1999) and Andreasen (2005), argue that racial categories do reflect biological realities, but nonetheless argue that such realities cannot support the kinds of discrimination that have historically been associated with them.[5]

I would like to propose that an understanding of the metaphysics of the senses shares some of the same features that make understanding species and race important. Parallel with the concept of species, the sensory modalities are

---

[4] A parallel discussion exists in biology and philosophy of biology over the metaphysics of the concept *gene*. As with the debate over the nature and reality of species, there is also debate over whether genes exist and, if so, what is their metaphysical nature, cf. (Beurton, Falk et al. 2000; Moss 2003; Fox Keller and Harel 2007).

[5] I could have spelled out much the same point about *gender categories* as I do about race here.

centrally important categories to any study of perception. Pick up any number of books about perception, say in sensory psychology or neuroscience or sensory anthropology, and you will find discussions of individual senses. It is not uncommon to find books on perception broken up into chapters, each focussing on a different sensory modality. Further, this division often is presented with little or no discussion of what grounds or justifies such a division. Such a division is taken literally for granted.

That lack of explicit justification for the division of perception into the categories of vision, audition, smell, etc. derives from the trait that the concept of sensory modalities shares with race (and gender): its ubiquity as a human category. While there is some disagreement over the specific senses posited,[6] the practice of dividing up the senses is apparently universal. As anthropologist Kathryn Linn Geurts (2002) puts it, "...a culture's sensory order is one of the first and most basic elements of *making ourselves human*. I define *sensory order* (or *sen-*

---

[6] For example, the Anlo-Ewe of Western Africa count a *balance* sense among the basic senses (see (Geurts 2002)). As sensory anthropologists Howes & Classen (1991) put it,

> Other cultures do not necessarily divide the sensorium as we do. The Hausa recognize two senses [citing Ritchie]; "the Javanese have five senses (seeing, hearing, talking, smelling and feeling), which do not coincide exactly with our five" [citing Dundes]. In short, there may be any number of "senses," including what we would classify as extrasensory perception—the "sixth sense." According to the Peruvian curer interviewed by Douglas Sharon in *Wizard of the Four Winds*, for example, a sixth clairvoyant sense opens up when all five other senses have been stimulated through the use of hallucinogens and other ritual elements.... Eduardo, the curer, describes this sixth sense as 'a 'vision' much more remote... in the sense that one can look at things that go beyond the ordinary or that have happened in the past or can happen in the future." (257-8)

Oddly, Nudds cites the book that this passage was taken from (Howes 1991) in support of his pronouncement that, "It is possible that some cultures distinguish fewer than five senses (by grouping together two senses we distinguish), but I have not been able to find a description of any culture that distinguishes more than five senses" (311).

sorium) as a pattern of relative importance and differential elaboration of the various senses, through which children learn to perceive and to experience the world and in which pattern they develop their abilities" (5, emphases in original). In other words, Geurts is observing that across human culture, sensory organization is one of the basic ways in which we enculturate our children and teach them who they are and how we all, as humans, interact with our world. Even if sensory categories do not have the powerful social implications that race and gender categories do, there is a value to studying other deeply held conceptual schemes, even if they do not lead to prejudice and injustice.[7]

In sum, the division of perception into different senses is centrally important to the study of perception and such a division is a ubiquitous and central human practice. Given this, it would behoove us to understand what kind of division this talk of the senses involves.

Further, if the discussion and arguments I present in this paper are correct, we (both we, the folk, and we, the investigators of sense) may be more confused about the nature of the senses and what kind of kind they are. If identifying (and, better, clearing up) confusion is a virtue, then I strive for that here.

III. Kinds of Kind

[7] I'm not convinced that sensory categories are as "innocent" as this implies. One need only look at social attitudes of "neurotypicals" towards those who are blind, deaf, etc. to see that there are likely to be important issues of social concern here. There is a growing body of literature in "disability studies" that is relevant to this point. I only wish to argue that *even if* sensory categories are innocent of such implications, they are nonetheless of interest relative to similarly entrenched categories with more apparent social implications. If they are *not* so innocent, then so much the better for my point here.

What are the possible answers to the question "what kinds of kind are the senses?" Here are a couple of possibilities:[8]

a) Senses as natural kinds: Some division of the senses (leaving open what that division is) constitute a set of natural or scientific kinds. That is, the division of perception into a number of senses is something that is *discovered* about the nature of the universe, not *invented* by humans; it is some kind of mind-independent metaphysical division of the universe. As Bird & Tobin (2010) put it: "To say that a kind is *natural* is to say that it corresponds to a grouping or ordering that does not depend on humans. We tend to assume that science is successful in revealing these kinds; it is a corollary of scientific realism that when all goes well the classifications and taxonomies employed by science correspond to the real kinds in nature. The existence of these real and independent kinds of things is held to justify our scientific inferences and practices" (emphasis in original). Therefore, this reading would say that the senses are kinds analogous to the way in which different chemical elements or, perhaps, fundamental subatomic particles are kinds. At one point, species were taken to be a paradigm case of natural kinds, but as noted earlier, Darwin upended that account.[9] I'll return to the senses as natural kinds in the final section.

---

[8] I do not want to claim that this list is exhaustive. I'm not even sure the options I present are mutually exclusive.

[9] Hacking (1991) offers a nice overview of the history of this term, tracing its origin back to J. S. Mill and John Venn in the late nineteenth century, although they were only giving a modern label to a concept with roots going back to at least Aristotle.

b) Senses as phenomenal kinds: In an oral response to an earlier paper of mine, Tom Polger once said, "Much of the bad press over qualia is well-deserved; but if there is one place experiential qualities have a safe home, I would've thought it would be with the sense modalities." He is not alone in his intuition; the natural place to talk about the phenomenal qualities of consciousness is the perceptual realm. Philosophers of mind like to speak of the sharp pain of a papercut, the tanginess of a lemon, the deep, velvety red of a rose.[10] The division of the senses into kinds could be the division of conscious perceptual experiences into different categories based on how it feels to experience them. This will be the topic of §V, below.

c) Senses as social kinds: While the senses clearly *can* be divided (we do so and have done so apparently since prior to the invention of written culture), such a division is conventional. Humans divide up the senses in response to cultural conditions. On this account, the senses are kinds analogous to the way that nonverbal gestures can be individuated. For example, in some Arabic cultures, sitting in such a way as to show the soles of one's feet to another is an offensive gesture, whereas Western Europeans might not even recognize sitting in this way as any kind of "gesture" at all. As noted above, this is Nudds' final position on the question and I'll return to this option at the end of §V, concerning phenomenal kinds, below.

---

[10] Although, as I describe in (Keeley 2009a), the strong association of qualia with *sensory* qualities is a mid-20th-century shift from earlier philosophical practice. This prior use reserved the term for *non*sensory or *multi*sensory phenomenal qualities, such as the feeling of effort or the quale of spaciousness.

d) Senses as functional kinds: Functional kinds are defined by the causal role they play in some larger system, rather than by any constitutive or phenomenal property that they might have. On a functional account, the senses are kinds analogous to the way that the different organ systems of the mammalian body can be individuated into systems: the respiratory system, the digestive system, or the circulatory system, or perhaps better: the way that different parts of any one of those systems can be divided up. It is important, however, to make sure that this use of function is firmly connected to the related-but-different concept of function as it is used in evolutionary theory. This is the topic of the next section, §IV.

IV. Senses as functional kinds

A functionalist is one who claims that psychological states are neither physical nor physiological states of a system but rather that they are Functional states, defined by their role within a causal description of some sort. Putnam (1967) introduces the notion of a functional Description by proposing that psychological systems can be described in relation to a Turing machine framework:

> A Description of *S* where *S* is a system, is any true statement to the effect that *S* possesses distinct states S_1, S_2, . . . S_n which are related to one another and to the motor outputs and sensory inputs by the transition probabilities given in such-and-such a Machine Table. The Machine Table mentioned in the Description will then be called the Functional Organization of *S* relative to that Description, and the S_i such that *S* is in state S_i at a given time will be called the Total State of *S* (at that time) relative to that Description. (226)

Cummins (1975) generalizes Putnam's account and reframes it in more general terms: "a function-ascribing statement explains the presence of the functionally characterized item *i* in a system *s* by pointing out that *i* is present in *s* because it has certain effects on *s*" (741). He notes that we find these kinds of explanatory strategies all the time in the description of artifacts, as when engineers produce schematic flowchart diagrams with symbols representing the different items; items described in terms of the functions they carry out (resistors, capacitors, etc.) (760). Cummins goes on to discuss how,

> Functional analysis in biology is essentially similar. The biologically significant capacities of an entire organism are explained by analyzing the organism into a number of "systems"—the circulatory system, the digestive system, the nervous system, etc.,—each of which has its characteristic capacities. These capacities are in turn analyzed into capacities of component organs and structures. Ideally, this strategy is pressed until pure physiology takes over, i.e., until the analyzing capacities are amenable to the subsumption strategy. We can easily imagine biologists expressing their analyses in a form analogous to the schematic diagrams of electrical engineering, with special symbols for pumps, pipes, filters, and so on. Indeed, analyses of even simple cognitive capacities are typically expressed in flow charts or programs, forms designed specifically to represent analyses of information processing capacities generally. (760-761)

So, in the case of the senses, we can use an approach like this to understand the nature of perception. As Cummins just described, after identifying the nervous system as one of the components of an organism, it is then further analyzed into its components, one of which would likely be a *sensory system*, alongside the *motor system*, as well as any number of systems situated between the "input" and "output" of the organism. Further, this sensory system would be fur-

ther analyzed into the different subsystems that we commonly think of as the different senses: a visual system, an auditory system, and so on. This is a not unreasonable way of capturing what some sensory scientists do in neuroscience and psychology.

It is important to keep in mind that these functional kinds are also functions in the sense of evolutionary biology; they are Darwinian functions that explain why organisms have evolved to have the traits that they have.[11] This characteristic of functional kinds is important for re-identifying those kinds in different evolutionary lineages, such as when biologists speak of the convergent evolution of vision in different taxa: Both vertebrates (e.g., humans) and mollusks (e.g., octopus) have evolved vision and possess eyes, but biologists believe that the most-recent common ancestor of vertebrates and mollusks had neither eyes nor vision. Given that humans and octopus eyes are physically different (as a result of their unrelated phylogenetic origins) what makes these structures both "eyes" is that they share an identifiable evolutionary function.

This is the sort of approach that Nudds (2011) has in mind when he speaks of "psychological kinds"[12]:

> My suggestion, then, is that the most plausible explanation of the distinction we make between senses is that we distinguish perceptions into perceptions of different senses on the basis of a reflective understanding of how those perceptions were produced. In doing so, we are distinguishing between perceptions produced by different kinds of sensory mechanism, and so our

[11] This is a point stressed by philosophers of biology, e.g., (Sober 1985; Kitcher 2003).

[12] Although, Nudds' terminology doesn't map cleanly onto mine. At times he uses the term "psychological kind" to refer to what I'm calling a "phenomenal kind," e.g., p. 336. Part of the confusion is that he sometimes talks of the putative function of the senses as producing certain phenomenal perceptual states.

> concepts of the senses must be concepts of different kinds of sensory mechanism. This provides an answer to the question of what constitutes a sense modality. A sense modality just is a kind of sensory mechanism, and all instances of, say, *seeing* something are instances of seeing that thing in virtue of their having been produced by a single kind of sensory mechanism—the sensory mechanism of vision. (314, emphasis in original)

Most of Nudds' paper is an exploration and eventual rejection of this approach. In particular, he understands this approach to require one-to-one mappings of sensory mechanisms onto the functional kinds, a requirement that he argues that contemporary perceptual psychology shows to be violated. Vision and audition (which he considers in some detail) have been shown to involve the operation of multiple perceptual mechanisms, such as the "dual stream hypothesis" of Milner and Goodale (1995; Goodale 1998), according to which there are separate pathways underlying the visual identification of objects (the "what" pathway) and the guidance of motor action (the "where" pathway). On the basis of this and the presence of similar features found by sensory psychologists in the other senses, Nudds concludes:

> That, I think, undermines the suggestion that the senses are natural kinds—it undermines the suggestion that the distinction we actually make between different senses tracks a natural distinction between kinds of psychological processes, and it shows that we cannot appeal to the psychological processes involved in perception to answer the question with which I began: What do all instances of seeing have in common in virtue of which they are instances of seeing? Whatever it is they have in common—whatever it is that makes a *visual* perception a *visual* perception—it is not that they are produced by a single kind of sensory mechanism. (335, emphases in original)

A few things should be noted here, in response. First, on Nudds' reading, functional kinds just are a subspecies of natural kind.[13] This makes sense if we take one of the key features of natural kinds to be that they are human-independent categories. The identification of sensory mechanisms within a functionalist frameworks seem appropriately mind-independent here; they are scientific discoveries, not inventions (if we are to adopt any reasonably realist account of science).

But, second, notice that his functionalist account is one that doesn't make much reference to the *evolutionary* aspects of function. As mentioned earlier, attributions of evolutionary function are useful in heading off the dead-end that Nudds finds himself in his deployment of sensory mechanisms. The sensory mechanisms of the octopus and human eyes are markedly different, but we can identify them both as functionally equivalent because of the role each organ plays in the lives/reproductive fitness of the organisms which possess them. As Kitcher (2003) puts it, "When we attribute functions to entities that make a causal contribution to complexes, there is, I suggest, always a source of design in the background. The constituents of a machine have functions because the machine, as a whole, is explicitly intended to do something. Similarly with organisms" (169). That background evolutionary context allows us to group together sensory

[13] He also briefly discusses and rejects a few other ways that the senses could be natural kinds. They might be anatomical kinds; that is, we might be able to differentiate the kinds by reference to the anatomical features of their respective sense organs (335-336). This view ultimately founders, he believes, on his claim that any anatomical account must ultimately presuppose a functionalist individuation of sensory mechanisms, an account that he has already shown to be wanting. He also discusses the option that I here describe as phenomenal kinds. In the end, he finds all these accounts wanting and is left with the remaining option that we count the senses we do because it is our social practice to do so, and nothing more.

mechanisms that too narrow a focus on the proximate causal analysis of mechanisms would classify as separate.

So, returning to the earlier point, if the senses are to be natural kinds in virtue of being functional kinds, what Nudds has shown us is that we cannot understand functional analysis here solely in proximate, psychological mechanism terms; instead, we need to understand functions more broadly in ultimate evolutionary terms, understanding not just the operation of the mechanisms, but understanding the role those mechanisms play in the evolutionary history of the organisms that possess them.[14]

V. Senses as phenomenal kinds.

One complication in all this questioning of the kind-hood of the senses is that the senses are not just biological categories, traits possessed by biological organisms. They are also phenomenal categories; that is, they are categories of conscious experience. As such, they lay at the foundation of a very deep way of dividing things up into kinds. They are qualities in the way of speaking where one says, "These two things differ not only quantitatively, but qualitatively." Or, "What we have here is not a difference in degree but a difference in kind."

This connection between this general sense of qualities and the elements of phenomenal experience is borne out by the importance of the concept *quale/*

---

[14] Two final notes to place this discussion in a larger context. First, I discuss the importance of such evolutionary (and also ontogenetic/developmental) considerations as the importance of understanding the "dedication" of putative sensory systems to particular modalities, see (Keeley 2002: 17-19), for a critical reply, see (Matthen 2012). Second, I should also note that there is some controversy over whether we should be realists with respect to evolutionary function. There are those, such as Dan Dennett, who reject realism about functional attributions. Perhaps unsurprisingly, I side with realists such as Kitcher. See the debate referenced in (Keeley 1999).

qualia to 20th century philosophy of mind. On the one hand, the concept of qualia is that they are, in some sense, the basic building blocks of phenomenal experience. Any given conscious experience will have a number of different qualia that make it up, including qualia of color and smell, as well as emotional tone, feelings of recognition or novelty, and the like. On the other hand, the term itself derives from talk of kinds. According to the *Oxford English Dictionary*, "quale" derives from the Latin *qualis* ("Of what kind") and the term means "The quality of a thing; a thing having certain qualities." So, quale stands as an important link between the idea of sensory qualities and the idea of kinds.[15]

There is some sense to saying that there is some kind of fundamental qualitative distinction between the experiences of different senses.[16] It is common to suppose that visual experiences are *just qualitatively different* from auditory experiences. This line of thought naturally gives rise to an understanding of the senses as natural kinds in the sense that the difference between the senses here is *given* to us, not invented by us. Metaphysically, the difference between the senses is as given as the difference between the chemical elements.

However, this account is problematic in a number of different ways. The crucial element of the account is that the basis for the difference between the

---

[15] I have previously discussed the early philosophical history of the concept "qualia" before. See Keeley (2009a).

[16] As Nudds (2011) observes, "Many philosophers suppose that there is an obvious answer to [the question of what differentiates the senses]. In order to perceive something one must have an experience of it. Seeing something requires having a *visual* experience of it, hearing something requires having an *auditory* experience of it, and so on. The different kinds of experiences involved in perceiving are what constitute perceiving with different senses. We *see* something just in case we perceive it in virtue of having a visual experience of it; *hear* something just in case we perceive it in virtue of having an auditory experience of it, and so on. To answer the question in this way is to give an *experiential account* of the senses." (312, emphases in original).

senses is given in experience. One can challenge the claim that a relevant datum is given.[17]

The issues here relate to concerns over the transparency of perception. In debates over the nature of phenomenal experience, there are those who argue that we can never fix on the nature of phenomenal experience *per se*, but that instead we always peer through the experience to that which is represented by the experience. As Block (2007) puts it, the idea of transparency is that, "...when I try to introspect my experience of the redness of a tomato, I only succeed in attending to the color of the tomato itself, and not to any mental feature of the experience. The representationalist thinks that we can exploit this intuition to explain phenomenal character in non-phenomenal terms" (611).

The experience itself is diaphanous, *transparent*. Or, so the transparency thesis holds. This alleged transparency of experience is taken by proponents of a representationalist account of consciousness as a "powerful motivation" for their view (Tye 2000: 45). According to strong versions of representationalism, the phenomenal character of experience is exhausted by the representational content of experience; there is nothing *more* to phenomenal experience beyond what is represented. This view is bolstered by the transparency thesis, in that if all we ever experience is perception is that which is represented in perception—be-

[17] One could also, of course, challenge the concept of givenness itself; one could argue that nothing is *ever* given in experience capabable of doing any interesting epistemic work. While I am sympathetic to such a line of argument, here I will restrict myself to the narrowerer claim that in this specific case, nothing is given in experience that can act as the basis of a way to distinguish the senses into kinds.

cause we "see through" the experience to the represented properties—then there is nothing "left over" requiring a *non-representational* explanation.[18]

My concern here is not debates concerning representationalist theories of consciousness.[19] Currently, there are philosophers of mind on both sides and the issue seems to be unsettled. Rather, the issue points to two significantly different ways in which the senses as phenomenal kinds might work. If one rejects representationalism, then one would hold that the experiences of different senses will differ intrinsically; that there is some "vision-y" character that all visual experiences share and which is experienced simply as different from the experiences of other senses.[20] On this account, visual experiences would wear their visual status on their sleeves, as it were. Put another way, the sensory modality of a perceptual experience would be *given* in that experience. This brute phenomenal difference between experiences would be the grounds for differentiating different phenomenal kinds; this would ground a literal qualitative distinction of kinds here. Let's call this view a Nonrepresentationalist Phenomenal Kinds view.

---

[18] I am greatly compressing a complicated argument here. See (Tye 2000: Ch. 3) for more. For some reasons to demur, see (Kind 2003).

[19] As an aside, in his 2011 paper, Nudds explicitly endorses the transparency thesis and denies that there is anything intrinsic to perceptual experiences themselves that can be used to differentiate the senses (cf., 312*ff*). In a footnote, he identifies this as a "a fundamental disagreement" with what I say in the opening sentences of my 2002 paper. For the record, I was taking no substantive stand in those introductory sentences. Indeed, I saw myself to be setting up the phenomenon or problem, much as Nudds himself does in the first paragraph of his own recent paper (which I quoted above on page 1).

[20] Perhaps because I haven't read enough on the non-representationalist side of this debate, I'm unaware of somebody making precisely this line of argument (specifically that *the sensory modality* of a perceptual experience is given in experience). Please let me know if you know somebody who makes this specific claim.

Representationalists (because of the transparency thesis) deny the existence of any such intrinsic character of experience beyond what is perceptually represented. Indeed, talk of some "vision-y" character of experience is precisely the kind of thing they are wont to deny. However, this is not to say that they do not talk of phenomenal kinds—their talk of perception is rife with such talk; they just ground it differently. For example, consider the following passages from arch-representationalist Michael Tye. He begins by reminding us of the representationalist view, taking vision as his example:

> Visual phenomenal qualities or visual qualia are supposedly qualities of which the subjects of visual experiences are directly aware via introspection. Tradition has it that these qualities are qualities of the experiences. Tradition is wrong. There are no such qualities *of experiences*. If we stipulate that something is a visual phenomenal quality or a quale only if it is a directly accessible quality of experience, then there are no visual phenomenal qualities or qualia. Still there are qualities of which the subjects of visual experiences are directly aware via introspection. They are qualities of external surfaces (and volumes and films), if they are qualities of anything. These qualities, by entering into the appropriate representational contents of visual experiences, contribute to the phenomenal character of the experiences. Thus, they may reasonably be called "phenomenal qualities" in a less restrictive sense of the term. (2000: 49, emphasis in original)

Tye then goes on to point out that such an analysis is not restricted to vision, but instead extends easily to the rest of perception:

> All of the above points generalize to other perceptual modalities. For example, we hear things by hearing the sounds they emit. These sounds are publicly accessible. They can be recorded. Similarly, we smell things by smelling the odors they give off. They, too, are publicly accessible. You and I can both smell the foul odor of the rotting garbage. Odors, like sounds, move through physical space. We taste things by tasting their tastes. One and the same taste can be tasted by different

> people. Some tastes are bitter, others are sweet. When we introspect our experiences of hearing, smelling, and tasting, the qualities of which we are directly aware are qualities we experience as being qualities of sounds, odors, and tastes. It seems very natural to suppose that among these qualities are the following: pitch, tone, loudness, pungency, muskiness, sweetness, saltiness, sourness. But this supposition is not needed by the argument. The crucial point again is that the qualities of which we are directly aware via introspection—*whatever* they turn out to be—are not qualities of the experience of hearing, smelling, and tasting. Rather, they are qualities of public surfaces, sounds, odors, tastes, and so forth, if they are qualities of anything at all (for, as before, the experiences may be hallucinatory). Change *these* qualities—the ones of which we can be directly aware via introspection—and, necessarily, the phenomenal character of the experience changes. Again, then, phenomenal character is best taken to be a matter of representational content. And again, there are no phenomenal qualities, conceived of as qualities of experiences. (2000: 49-50, emphasis in original)[21]

In these passages, it is clear that Tye is distinguishing between the phenomenal qualities of the different senses. The entire discussion is organized in relation to the different senses. But what exactly grounds these distinctions? The answer is revealed when Tye responds to counter examples raised by Ned Block (1995; 1996) and others that there are cross-modal cases where we have *different* phenomenal experiences despite those experiences having the *same* representational content, as when we come to know that there is a round surface in front of me either by *seeing* that surface or by reaching out and *feeling* it with my hands. In this case, there is a clear phenomenal difference (seeing vs. tactually feeling a surface) despite the same representational content (that there is a round surface 15 centimeters in front of my nose).

---

[21] In the passages that follow these, Tye similarly extends his account to cover the rest of phenomenal experience, including bodily sensations (including pain), moods and emotions.

In responding to this kind of counter example, Tye has to lay out how a representationalist distinguishes phenomenal kinds of one modality (visual) from another (tactile):

> One obvious immediate reply the representationalist can make is that in seeing the shape, one has an experience as of color. But color isn't represented in the content of the haptic experience. Conversely, temperature is represented in haptic experience but not in the visual one (or at least not to the same extent). Likewise, there is much more detailed representation of degree of solidity in the haptic experience. Another representational difference pertains to the location of the shape. In vision, the shape is automatically represented as having a certain two-dimensional location relative to the eyes. It is also normally represented as being at a certain distance away from the body. In the haptic case, however, shape is represented via more basic touch and pressure representations of contours derived from sensors in the skin.[22] Here the shape is represented as belonging to a surface with which one is in bodily contact. Moreover (and relatedly) in the haptic experience, there is no representation of the shape's two-dimensional location relative to the eyes. Finally, and very importantly, in the visual case, there is representation not only of viewer-independent shape but also of viewer-relative shape (e.g., being elliptical from here). The latter property, of course, is not represented in the haptic experience.[23] (95)

This then allows us to spell out what I will call the Representationalist Phenomenal Kinds view. What Tye seems to be proposing here is that we can

---

[22] I have no idea why Tye thinks these are "more basic" or even what this means in this context.

[23] As with the previous footnote, I also find this last point somewhat baffling. In claiming that the haptic sense does not represent the body-relative position of the surface, Tye must be distinguishing the contents of proprioception from those of pressure sensation. That is, if I feel my fingers pressed up against a surface 15cm in front of my face, the pressure sensors themselves do not give rise to that representation of distance; that distance representation is a product of the simultaneous proprioceptive sense of the positions of my limbs at that moment (elbow and wrist bent just this much, etc.) Fine, but it is odd to segregate pressure and proprioception sensation while simultaneously lumping pressure and temperature sensation together as a single sense of "touch" despite the fact that these are carried out by different peripheral sensory systems in the skin. At least, without further discussion, these distinctions seem arbitrary, especially given that the sense of touch is famously the most problematic for a proper objects account of the senses (see (Keeley 2002; 2009b)).

distinguish between the different senses on the basis of what Aristotle called the "proper" or "special objects" of perception.[24] The proper object(s) of a sense are those qualities that *only* that sense can elicit; for example, we only come to experience color by vision, only come to experience temperature through touch. This exclusive connection between each sensory modality and some particular proper object experienced as a result of its action—color for vision, temperature/pressure for touch, odor for smell, flavors for taste, pitch/loudness for hearing—gives us a representationalist means for dividing up the senses into distinct categories. And circling back to the points made at the beginning of this section, pitch, flavor, color, etc. are paradigm examples of "qualities" at play when we talk about differences of quality vs. difference of quantity. The difference between a whisper and a nearby crash of thunder is a difference of *quantity* (of loudness) whereas the difference between a whisper and a nearby flash of lightning is a difference of *quality*.

A phenomenal kinds account of the senses—in either the Representationalist or Nonrepresentationalist version—is interesting and has a degree of plausibility in the way that it matches up with our own perceptual experience. In the end, it may turn out to offer a coherent and empirically-adequate account of the division of the senses into kinds. However, I have my doubts.

First, I don't understand how such an account can make sense of senses that lack any phenomenal character, such as a putative "pheromone (or vomero-

[24] I discuss this Aristotlean approach at length in (Keeley 2002; 2009b).

nasal) sense" in humans.[25] Some scientists report a phenomenon akin to the once-common philosophical example of "chicken sexing" whereby subjects can reliably make behavioral discriminations of vomeronasal stimuli, but these same subjects report no phenomenal differences in their experiences. They feel as though they are guessing. Proponents of phenomenal kinds can simply (and consistently) deny that this putative sense actually *is* a sense; indeed, they seem to be forced to. OK, but what of the senses of non-human animals; especially those where the presence of a vomeronsal sense is well established, and is studied alongside other senses? How exactly do we make the sense/nonsense distinction in nonhuman animals; that is, how do we tell when they are "guessing" or acting unconsciously? Again, it would be consistent for believers in phenomenal kind accounts of the senses to deny that animals have senses, but that seems to be a more draconian move, to say the least.

A second problem for phenomenalist accounts is that recent work in sensory psychology is increasingly undermining the empirical viability of the proposed connection between senses and their unique qualities. In other words, scientists are increasingly showing that, in practice, the phenomenal qualites of experiences we have are the product of multiple senses, not just one. So, in the McGurk effect, the sound that you hear is a product of *both* what your ears hear *and* what your eyes see. Change (only) what the eyes see (in this case, what an

---

[25] I raise this example in (Keeley 2002: 23ff). Noë (2004: 107-111) and I (Keeley 2009b: 231-238) have debated issues related to this. Although Noë explicitly attempts to avoid endorsing a phenomenalist account, I argue that he nonetheless ends up running into the same problems that somebody such as Tye will need to overcome.

interlocutors lips look like) and your auditory experience changes.[26] Similar results have been found for other sensory combinations: what you hear effects what things taste like, what you hear effects what things look like, etc.[27] This is important for the representationalists because, such cases threaten to show that the representational feature they need to be uniquely connected to a given sense (if it is to play the role of identifying the phenomenal kind) is not, in fact, unique.

For the nonrepresentationalist, these results from sensory psychology are threatening in a slightly different way, which brings me to my third concern with phenomenalist accounts of the senses. In arguing that the senses can be differentiated into kinds by the presence of some given phenomenal quality, such accounts ignore the potential theory-ladenness of introspection. What the theory-ladenness of introspection means is that what we experience (and therefore what we can claim to introspect) is, in part, a function of the theoretical categories we bring to that introspection.[28] If this is the case, these experienced qualities are not *given*, in that if we had different categories and a different understanding of

---

[26] Of course, calling it an "auditory experience" begs the question here. To the extent that what people report hearing is a product of what effected their eyes, then on the representationalist account, the experience would not be "auditory" but rather some mixture of the two. If there are no unique features represented by individual senses, then the representationalist account will be unable to differentiate perceptions into different sensory kinds. This is my point.

[27] See (Calvert, Spence et al. 2004; Spence and Driver 2004) for more on multisensory perception. Also, Nudds (2011: 335) also discusses the importance of multisensory perception for accounts of the senses.

[28] In this way, the theory-ladenness of introspection is intended to parallel the thesis of the theory-ladenness *of perception*, familiar from discussions in the philosophy of science, see (Hanson 1958). The idea is that what one observes is, in part, a function of the theory one brings to an observation. For example, in a real sense, what a heliocentrist and a geocentrist observe when peering eastward at dawn is different and is a function of those theories. I take the notion of the theory-ladenness *of introspection* from (Churchland 1985), but he credits ideas found in (Feyerabend 1963).

how the action of our senses gives rise to our perceptual experiences, we experience perception differently. Further, as indicated by the anthropology of senses (see footnote 6 above), humans do, in fact, have different categories and different understandings of the nature and number of senses.

If the theory ladenness of introspection is true, then it would imply that, in essence, the senses as phenomenal kinds are, in fact, in part, a product of social categories. That is, phenomenal kinds would be, in part, reducible to social kinds, in that how one is raised and enculturated would provide one with the categories of sense and these, in turn, would play an important role in how one phenomenally experiences the process of perception. At least, the possibility of the theory ladenness of introspection poses some interesting lines of investigation for sensory anthropologists—a rather young sub-discipline of anthropology—to explore. Does counting balance among one's basic categories of the senses change how one reports experiencing perception?

However, these three worries are just that: *worries*. They involve more open questions to those that want to defend and explicate a phenomenalist account of the senses than refutations of this approach.

IV. Implications and Conclusions

In this paper, I have considered a number of different ways of thinking of the senses as kinds. OK, but which way is the *correct* way? Ultimately, are the senses natural kinds? And, if so, specifically functional kinds? Phenomenal kinds? Social kinds? Some other notion of kinds? My response is to resists the

implication of such questions by resisting the implication that there has to be a *single* answer to the question of what kinds of kind are the senses. I propose that we embrace a form of pluralism with respect to the senses. To see why I say this, however, let me take a quick detour in what might seem to be the opposite direction: the notion that the senses are not any *kind* at all; that the application of "kind talk" to the senses is just a bit confused from the get go.

Consider the following explanation of what a natural kind is supposed to be, taken from the *Concise Routledge encyclopedia of philosophy*: "Objects belonging to a natural kind form a group of objects which have some theoretically important property in common. ... Natural kinds are contrasted with arbitrary groups of objects such as the contents of dustbins, or collections of jewels. The latter have no theoretically important property in common: They have no unifying feature" (Daly 2000: 612-613). Notice that this account of natural kinds has two important features. First, there is the *collection of entities* brought together under the natural kind description ("a group of objects"). Second, there is the *property* that so defines that collection (the "theoretically important property," the "unifying feature"). That is, when we normally speak of natural kinds, there are two components, reflected in the term itself: There are the *kinds* (the collection of entities) sorted according to some property (in this case, a nonhuman, *natural* property. This, in contrast to, say, artificial kinds which is a collection of things delineated by some perhaps arbitrary, human property).

Noticing this reveals that there is something deeply odd about much of the discussion I have engaged in above. Unlike what is the case when considering

situations such as species, genes, races and genders, the sensory examples I have been discussing actually sound much more like the properties that define the kinds and not the kinds themselves. That is, when talking about the metaphysical nature of vision as versus hearing, for example, that nature is closer to talk of the thing by which we define a collection of entities, not the entities themselves. Consider the senses as functional kinds: Vision is a function (a property) that allows one to class certain mollusks and vertebrates (a collection of entities) together into the same category. Even on the less evolutionary reading of functions that Nudds discusses, a sense is identified with a particular perceptual mechanism; possessing that mechanism is something either organisms have or do not. It is a *state* that an entity can (or cannot) be in. Again, the sense is a property not an entity. Or, consider the case of phenomenal kinds: Here it is even more explicit that what we are dealing with are properties (the possession of phenomenal qualities) instead of collections of entities. In all of these discussions, we have not been careful enough to distinguish the properties (e.g., having vision) from entities (e.g., the sighted). The latter are the "kinds"; the former are the properties that define the kinds.

Recognizing that senses are more properly thought of as the defining properties of groups of kinds rather than the kinds (as collections of entities) themselves in turn supports a kind of pluralism in relation to talk of the senses. It is a commonplace to identify a large number of properties possessed by any given entity and it is equally commonplace to cross-categorize the kind-groupings

into which we place a given entity depending on what properties one is attending to.

Further, as the Daly quotation above also stresses, the properties that are important in natural kinds are "theoretically important" properties. It is far from clear that there is only *one* theory—and, hence, only one set of theoretically important properties—in the study of the senses. I take it that my discussion of functional and phenomenal approaches to the senses demonstrates that. The concerns of psychophysicists and philosophers of perception concerned with the understanding phenomenal character of perceptual experience are different from neuroethologists, comparative biologists, and those interested in the evolution of sensory systems. This plurality of theoretical interest begets a plurality of kind-talk. Add in the "folk" (who, according to previously cited anthropologists, do not universally share intuitions about the number and identity of the senses) and you get even more ways of speaking about the senses.[29]

At the end of the day, I am not confident that I have answered my title's question to anybody's satisfaction. I plan to think about it further myself and look forward to your own thoughts. But, I hope I have demonstrated the complexity of the issues involved and will spur you to come up with your own answer.

---

[29] As Haddock and Dupré (2006) put it in their own encyclopedia entry on "Natural Kinds", "This possibility of diverging intentions suggests that one kind term might be a natural-kind term among a group of scientists (given how they use it) and a functional-kind term among a group of lay persons (given how they use it)" (505). (In their terminology, a "functional kind" is what I've been calling a "social kind" here.)

Bibliography

Andreasen, R. O. (2005). "The meaning of 'race': Folk conceptions and the new biology of race." The Journal of Philosophy **102**(2): 94-106.

Appiah, K. A. (1996). "Race, culture, identity: Misunderstood connections." Color conscious: The political morality of race. K. A. Appiah and A. Guttmann, Eds. Princeton, NJ, Princeton University Press.

Beurton, P., R. Falk and H.-J. Rheinberger, Eds. (2000). The concept of the gene in development and evolution. Historical and epistemological perspectives. Cambridge, Cambridge University Press.

Bird, A. and E. Tobin. (2010). "Natural kinds." from <http://plato.stanford.edu/archives/sum2010/entries/natural-kinds/>.

Block, N. (1995). "On a confusion about a function of consciousness." Behavioral and Brain Sciences **18**(2): 227-287. Reprinted in *The Nature of Consciousness: Philosophical Debates*, eds. Ned Block, Owen Flanagan, and Güven Güzeldere. Cambridge, Mass.: The MIT Press, 1997, 375-415.

Block, N. (1996). "Mental paint and mental latex." Philosophical issues: Perception. E. Villanueva, Ed. Atascadero, Cal., Ridgeview Publishing Company. **7:** 19-49.

Block, N. (2007). "Bodily sensations as an obstacle for representationalism." Consciousness, function and representation: Collected papers, volume 1. N. Block, Ed. Cambridge, MA, The MIT Press (A Bradford Book)**:** 611-616.

Boyd, R. (1999). "Homeostasis, species, and higher taxa." Species: New interdisciplinary essays. R. A. Wilson, Ed. Cambridge, MA, The MIT Press (A Bradford Book)**:** 141–185.

Calvert, G., C. Spence and B. E. Stein, Eds. (2004). The handbook of multisensory processes. Cambridge, MA, The MIT Press (A Bradford Book).

Churchland, P. M. (1985). "Reduction, qualia and the direct introspection of brain states." The Journal of Philosophy **82**(1): 8-28. Reprinted as Chapter 3 of Churchland, P.M. *A Neurocomputational Perspective: The nature of mind and the structure of science.* Cambridge, MA: The MIT Press (A Bradford Book): 47-66.

Cummins, R. (1975). "Functional analysis." Journal of Philosophy **72**: 741-765. Excerpts reprinted in Ned Block (ed.), *Readings in Philosophy of Psychology*, Cambridge, MA: MIT Press.

Daly, C. (2000). "Natural kinds." Concise Routledge encyclopedia of philosophy. R. Staff, Ed. New York, Routledge**:** 612-613.

Feyerabend, P. K. (1963). "Materialism and the mind-body problem." Review of Metaphysics **17**.

Fox Keller, E. and D. Harel (2007). "Beyond the gene." PLoS ONE **2**(11): e1231.

Geurts, K. L. (2002). Culture and the senses: Bodily ways of knowing in an african community, Berkeley, CA, University of California Press.

Ghiselin, M. (1974). "A radical solution to the species problem." Systematic Zoology **23**: 536–544.

Goodale, M. A. (1998). "Visuomotor control: Where does vision end and action begin?" Curr Biol **8**(14): R489-491.

Griffiths, P. E. (1999). "Squaring the circle: Natural kinds with historical essences." Species: New interdisciplinary essays. R. A. Wilson, Ed. Cambridge, MA, MIT Press**:** 208–228.

Hacking, I. (1991). "A tradition of natural kinds." Philosophical Studies **61**(1/2): 109-126.

Haddock, A. and J. Dupré (2006). "Natural kinds." Encyclopedia of philosophy. D. M. Borchert, Ed. Detroit, Macmillan Reference USA. **6:** 503-505.

Hanson, N. R. (1958). Patterns of discovery: An inquiry into the conceptual foundations of science, Cambridge, UK, Cambridge University Press.

Howes, D., Ed. (1991). The varieties of sensory experience: A sourcebook in the anthropology of the senses. Toronto, University of Toronto Press.

Howes, D. and C. Classen (1991). "Conclusion: Sounding sensory profiles." The varieties of sensory experience: A sourcebook in the anthropology of the senses. D. Howes, Ed. Toronto, University of Toronto Press**:** 257-288.

Hull, D. (1978). "A matter of individuality." Philosophy of Science **45**: 335–360.

Keeley, B. L. (1999). "Fixing content and function in neurobiological systems: The neuroethology of electroreception." Biology & Philosophy **14**: 395-430.

Keeley, B. L. (2002). "Making sense of the senses: Individuating modalities in humans and other animals." The Journal of Philosophy **99**(1): 5-28. Reprinted in Macpherson, F., Ed. (2011). *The senses: Classic and contemporary philosophical perspectives*. Oxford, Oxford University Press, 220-240.

Keeley, B. L. (2009a). "The early history of the quale and its relation to the senses." Routledge companion to philosophy of psychology. J. Symons and P. Calvo, Eds. London, Routledge**:** 71-89.

Keeley, B. L. (2009b). "The role of neurobiology in differentiating the senses." Oxford handbook of philosophy and neuroscience. J. Bickle, Ed., Oxford University Press.

Kind, A. (2003). "What's so transparent about tranparency?" Philosophical Studies **115**: 225-244.

Kitcher, P. (1984). "Species." Philosophy of Science **51**: 308–333. Reprinted in Philip Kitcher, *In Mendel's mirror: Philosophical reflections on biology*, New York: Oxford University Press, 2003: 113-134.

Kitcher, P. (1989). "Some puzzles about species." What the philosophy of biology is: Essays dedicated to David Hull. M. Ruse, Ed. Dordrecht, Kluwer Academic Publishers**:** 183-208. Reprinted in Philip Kitcher, *In Mendel's mirror: Philosophical reflections on biology*, New York: Oxford University Press, 2003: 135-158.

Kitcher, P. (1999). "Race, ethnicity, biology, culture." Racism. L. Harris, Ed. Amherst, NY, Humanity Books**:** 87-117. Reprinted in Philip Kitcher, *In Mendel's mirror: Philosophical reflections on biology*, New York: Oxford University Press, 2003: 230-257.

Kitcher, P. (2003). "Function and design." In Mendel's mirror: Philosophical reflections on biology. P. Kitcher, Ed. New York, Oxford University Press**:** 159-176.

Matthen, M. (2012). The individuation of the senses.

Milner, A. D. and M. A. Goodale (1995). The visual brain in action., New York, Oxford University Press.

Moss, L. (2003). What genes can't do, Cambridge, MA, The MIT Press.

Noë, A. (2004). Action in perception, Cambridge, MA, The MIT Press.

Nudds, M. (2004). "The significance of the senses." Proceedings of the Aristotelian Society **CIV**(1): 31-51.

Nudds, M. (2011). "The senses as psychological kinds." The senses: Classical and contemporary philosophical perspectives. F. Macpherson, Ed. Oxford, Oxford University Press**:** 311-340.

Putnam, H. (1967). "The nature of mental states." Art. Mind. And religion. W. H. Capitan and D. O. Merrill, Eds. Pittsburgh, Pittsburgh University Press**:** 37-48. Reprinted in Rosenthal (1971:150-161).

Sober, E. (1985). "Panglossian functionalism and the philosophy of mind." Synthese **64**(2): 165-193.

Spence, C. and J. Driver (2004). Crossmodal space and crossmodal attention, New York, Oxford University Press.

Stanford, P. K. (1995). "For pluralism and against realism about species." Philosophy of Science **62**(1): 70-91.

Tye, M. (2000). Consciousness, color, and content, Cambridge, MA, The MIT Press (A Bradford Book).

Wilson, R. A., Ed. (1999). Species: New interdisciplinary essays. Cambridge, MA, The MIT Press (A Bradford Book).

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

Newton on Matter and Space in *De gravitatione et aequipondio fluidorum*

Hylarie Kochiras

Abstract
This paper explicates the concepts of matter and space that Newton develops in *De gravitatione*. As I interpret Newton's account of created substances, bodies are constructed from qualities alone, as configured by God. Although regions of space and then "determined quantities of extension" appear to replace the Aristotelian substrate by functioning as property-bearers, they actually serve only as logical subjects. An implication of the interpretation I develop is that only space is extended by having parts outside parts; material bodies are spatially extended only in a derivative sense, via the presence of their constitutive qualities or powers in space.

Newton develops his account of material body in what Howard Stein has called the "creation" story or hypothesis. This account has also been called the "determined quantities of extension hypothesis" (Slowik, 2009), since Newton marks the account as speculative and develops it by associating various conditions with "determined quantities of extension".[1] I shall follow Stein's terminology, however, for reasons concerning Newton's account of minds, as explained later.[2] Understanding the account of body depends upon properly understanding these determined quantities of extension and their relation to space (extension) itself. It is therefore important briefly to review *De gravitatione*'s claims about space.

Features of space

For Newton, space is an existence condition for any substance and "an affection of every kind of being".[3] This latter description refers to the manner of existing in nature, a manner of existing quite different from that of an abstract entity or a number, as J.E. McGuire has

[1] See *De gravitatione* in *Isaac Newton: Philosophical Writings*, 27: "I am reluctant to say positively what the nature of bodies is, but I would rather describe a certain kind of being similar in every way to bodies..."; and 28: "And hence these beings will either be bodies, or very similar to bodies. If they are bodies, then we can define bodies as *determined quantities of extension which omnipresent God endows with certain conditions."*
[2] See Stein, "Newton's Metaphysics", 275. Slowik refers to that account of bodies as the "Determined Quantities of Extension" or "DQE" hypothesis (see "Newton's Metaphysics of Space", 2009, 438.) I follow Stein's terminology in part to avoid reifying the quantities of extension, and in part for a reason concerning minds, as discussed at the end of §4.
[3] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

explained.[4] As space is an affection of every kind of being, so is it a condition for their existence. As Newton asserts in a well known remark, one repudiating the concept of spirits as transcendent, "No being exists or can exist which is not related to space in some way. God is everywhere, created minds are somewhere, and body is in the space that it occupies; and whatever is neither everywhere nor anywhere does not exist."[5]

Since space is an existence condition of substances, it is not surprising that Newton takes it to have its own manner of existing. It is neither substance, he emphasizes, nor accident.[6] That it is not an accident inhering in a subject means, in part, that as an affection of every kind of being, it cannot be localized to any one being. Accordingly, it is independent of bodies; if all bodies were annihilated, it would continue to exist unchanged.[7] Space more nearly resembles a substance than an accident, Newton indicates, and as we shall see later, he ascribes a degree of "substantial reality" to it. Indeed, he cites it as the one thing that can in some circumstances be conceived apart from God—a feature he will use to attack Descartes' account of matter as atheistic.[8] Yet though it has some substantial reality, still space is not a substance. For one thing, it is "not absolute in itself, but is as it were an emanative effect of God."[9] Its not being

[4] Pointing to the manuscript 'Tempus et Locus' (c. 1692-93), as providing "Newton's most succinct statement of how place and time relate to existing things". McGuire explicates that statement as follows: "Newton answers the question: what is it for anything to exist in nature? It is to exist in a place and at a time. As the text implies, existing in place and time is what counts as actually existing, in contrast, for example, to existing in the manner of an abstract entity or as a number. This contention is supported by Newton's use of the phrase 'rerum natura'....." (McGuire, "Existence, Actuality and Necessity: Newton on Space and time", 465)

[5] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 25.

[6] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21-22. The ultimate source of Newton's view that space is neither substance nor accident is Renaissance thinker Francesco Patrizi da Cherso (1529-1597). Patrizi additionally held space to be wholly distinct from body, indeed a condition for matter's existence, and to be immutable, indivisible, and immobile. See F. Patrizi, 'On Physical Space' (De Spacio Physico), translated and commentary by B. Brickman, Journal of the History of Ideas, 4:2 (1943), especially 224–245. As Edward Grant explains (*Much Ado about Nothing*, 206-207), Patrizi is also the source of a surprising explanatory remark following Newton's claim that space has distinguishable parts, whose common boundaries may be called surfaces. Newton then goes on to explain that in space there are "there are everywhere all kinds of figures, everywhere spheres, cubes, triangles, straight lines, everywhere circular, elliptical, parabolical, and all other kinds of figures, and those of all shapes and sizes, even though they are not disclosed to sight....so that what was formerly insensible in space now appears before the senses....We firmly believe the space was spherical before the sphere occupied it, so that it could contain the sphere....And so of other figures." (Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21-22).

[7] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 22. See also 21: as "an affection of every kind of being", it is not a "proper affection" which is to say an action.

[8] See Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 31: "If we say with Descartes that extension is body, do we not manifestly offer a path to atheism, both because extension is not created, but has existed eternally, and because we have an idea of it without any relation to God, and so in some circumstances it would be possible for us to conceive of extension while supposing God not to exist?" On space's inability to produce effects, see *Newton: Philosophical Writings*, p 21-22, 34.

[9] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21. That space is not a substance cannot fully be explained by its dependence upon God, in virtue of being an emanative effect of God. For as will be emphasized later, Newton accepts not only the strong sense of substance but also the weak sense, which applies to things dependent upon God, in particular, created minds and bodies. Although I cannot here address the question of how Newton understands an emanative effect, I am sympathetic to McGuire's view that the relation of space to God is one of "ontic dependence". (See McGuire, "Existence, Actuality and Necessity: Newton on Space and time", 480: "the relation between the existence of being and that of space is not causal, but one of ontic dependence".) McGuire's view provides an alternative to the three that Gorham (September, 2011) identifies as

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

absolute could not by itself explain why it is not a substance; for neither are created substances absolute in themselves, being dependent upon God. Yet created substances have a different relation to God, precisely in virtue of having been created. There is also another important difference. Substances act, whereas space produces no effects.[10]

Though neither substance nor attribute, space is not nothing, Newton emphasizes, for it has properties. The properties he describes indicate a Euclidean space, three-dimensional, homogeneous, and infinite. Space is also eternal and immutable, and though parts may be distinguished within it, those parts are motionless and indivisible.[11] It is these features—the immobility and indivisibility of space's distinguishable parts—that are especially significant for Newton's account of body.

The creation hypothesis and the definition of body

Newton develops his creation hypothesis in two stages, first ignoring mobility but subsequently introducing it. He begins from the realization that we can temporarily make regions of space impervious to other bodies by moving our own bodies into them, observing that this might somehow simulate the divine power of creation. By his will alone, God "can prevent a body from penetrating any space defined by certain limits".[12] Such an entity would either be a body, or would be indistinguishable from bodies by us.[13] For if God made some region above the earth impervious to bodies and all "impinging things", it would be like a mountain; it would reflect all impinging things, including light and air, and it therefore would be visible and colored, and would resonate if struck.[14]

These entities would be very similar to corporeal particles, Newton notes, except for this important feature: he has imagined them to be motionless. For an entity to be a body, or at least

'Independence', 'Causation', and 'Assimilation'. Gorham defends Assimilation, arguing that space and time are attributes of God, and indeed identical to God (and thus to one another); see Gorham, September, 2011, especially 289-92 and 298-304.

[10] As I argue in §4, Newton takes God to be identical to his attributes, and fundamental to his creative power, that is, omnipotence; yet in doing so Newton does not eliminate substance but rather gives a reductive account of it. I note here that I reject the interpretation recently advanced by Geoffrey Gorham, though his arguments are intriguing. According to Gorham, God is identical to his attributes, but his attributes include space and time, and hence he is identical to space and time. (See Gorham, September, 2011, especially 289-92 and 298-304). In §4, I indicate the difficulties I see with that view.

[11] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 22, 25, 26.

[12] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 27.

[13] Newton means to emphasize that we cannot know matter's "essential and metaphysical constitution" (*De Gravitatione,* in *Newton: Philosophical Writings*, 27), or indeed the essence of any substance. This conviction reappears in later writings, including the General Scholium, where he writes, "We certainly do not know what is the substance of any thing. We see only the shapes and colors of bodies, we hear only their sounds, we touch only their external surfaces….But there is no direct sense and there are no indirect reflected actions by which we know innermost substances."(*Principia*, 942.) In this respect his account of body is strongly empirical.

[14] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

to resemble bodies in all humanly perceptible ways, it must be mobile. He therefore now adds that the hypothesized entities are capable of being moved from place to place, and in a law-governed way, a feature that is relatively new to conceptions of body.[15] Additionally, the entities can stimulate perceptions in minds and be operated upon by minds.[16] The hypothesized entities are now just like bodies, being perceptible, and having shape, tangibility, mobility, and the ability both to reflect and be reflected. They therefore could be "part of the structure of things", just like "any other corpuscle".[17] This enables Newton to provide a definition of body (insofar as we can know them).

> We can define bodies as *determined quantities of extension which omnipresent God endows with certain conditions.* These conditions are: (1) that they be mobile, and therefore I did not say that they are numerical parts of space which are absolutely immobile, but only definite quantities which may be transferred from space to space; (2) that two of this kind cannot coincide anywhere, that is, that they may be impenetrable, and hence that oppositions obstruct their mutual motions and they are reflected in accord with certain laws; (3) that they can excite various perceptions of the senses and the imagination in created minds, and conversely be moved by them, which is not surprising since the description of their origin is founded on this.[18]

One of the interesting things about this definition is that Newton sees it as serving theological goals, as will become evident from his commentary, and yet it is firmly rooted in experience. The fundamental features of our experience with bodies appear in the definition: their mobility; the mutual impenetrability that results in law-governed reflections of other bodies, light, and air; and the sensations they produce in us, such as those of color. Newton's remark at the end of the passage highlights the fact that experiences, specifically perceptions, make his description of the bodies' origin possible. For if bodies lacked the power to produce sensations, we could never have any ideas of them.[19] It is notable that Newton specifies condition (3), the

[15] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28. In an otherwise quite different thought experiment, which appears in *Le Monde*, Descartes imagines bodies that move "in accordance with the ordinary laws of nature"; see CSM 1, 90. Of interest here is Katherine Brading's article "On Composite Systems: Descartes, Newton, and the Law-Constitutive Approach" (2011).

[16] "For it is certain that God can stimulate, our perception by means of his own will, and thence apply such power to the effects of his will." (Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28)

[17] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

[18] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28-29. A definition given in 1678 by Robert Hooke contains some intriguing similarities. After asserting that the universe consists in body and motion, he writes, "I do therefore define a sensible Body to be a determinate Space or Extension defended from being penetrated by another, by a power from within." He also speculates that body and motion might ultimately be "one and the same". See Hooke, Lectures *Potentiae Restitutiva*, or of Spring, Explaining the Power of Springing Bodies, 1678, 338-340. How near the similarity really is, however, is a question I will not pursue here.

[19] Geoffrey Gorham interprets this remark very differently. On his view, Newton's remark that the description of bodies' origin is founded upon sensations indicates that he takes the capacity to produce sensations to be both necessary and sufficient for body-hood. In connection with that claim, Gorham argues that Newton ultimately sees his conditions of mobility and impenetrability

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

power to produce sensations, as distinct from condition (2), impenetrability. One reason for distinguishing them is that in the hypothesis' context, the first creation of matter, impenetrability could not be sufficient to produce sensations in minds. For if any minds existed when God first created matter, no human bodies would exist to touch it, and so the mutual impenetrability of bodies could not then produce sensations in minds. Yet there is another explanation for including condition (3) as independent of condition (2): even in the context of actual experiences, Newton does not seem to consider sensations as explicable solely in terms of impenetrability. He rather seems to share a belief common in the early modern period—that while the contact of light particles with the eye and food particles with the tongue seem to play some necessary role, they are not sufficient for the production of sensation, and so some role must be attributed to God.[20]

The definition's third condition is thus the basis for Newton's claim that Descartes' account of matter leads to atheism, while his own confirms God's existence. As indicated above, he takes space to be the one thing sometimes conceivable apart from God, since it produces no sensations or other effects, and so by identifying matter with extension (space), Descartes allows that matter is conceivable apart from God.[21] For as Newton indicates elsewhere, "we find almost no other reason for atheism than this notion of bodies having, as it were, a complete, absolute and independent reality in themselves."[22] On his own account, bodies are not conceivable apart

as superfluous; these "do no independent work of their own"(Gorham, Jan.2011, 24). I contest Gorham's conclusion about those conditions in §2.5.

[20] Here I disagree with Geoffrey Gorham, who argues that Newton actually intends his third condition, the capacity to produce sensations in minds, to resolve a problem about distinguishability (a problem that has concerned several commentators but did not, in my view, concern Newton, for reasons I indicate later in this section). On Gorham's view, if Newton did not intend his third condition to resolve that problem, it would be superfluous: "If the DQE's are impenetrable, they will be solid to touch, reflect light, perturb the air when struck, and so on. Since these are the means by which the senses perceive familiar bodies, why the need for God to affix also the special power to produce sensations? The answer seems to be that impenetrability alone is inadequate to distinguish bodies from the unfavored portions of absolute space." (Gorham, January 2011, 23). Yet as I have argued, Newton does not see the production of sensation as reducible to impenetrability, either in the context of matter's first creation, when no human bodies would exist even if minds did, or in his actual context, in which human bodies do exist. He takes a line similar to that found in Locke's *Essay*. Despairing of the ability of the mechanical hypothesis to reduce sensations to the shapes, sizes, and motions of particles, Locke suggests that the production of sensations must be attributed to God. Or, on an interpretation associated with Ayers, Locke thinks that we invoke superaddition because our powers of understanding are too limited to grasp how God might have enabled matter to produce sensations; my thanks to James Hill for discussion of the point.

[21] "If we say with Descartes that extension is body, do we not manifestly offer a path to atheism, both because extension is not created but has existed eternally, and because we have an idea of it without any relation to God, and so in some circumstances it would be possible for us to conceive of extension while supposing God not to exist?" (*De Gravitatione*, *Philosophical Writings*, 31). Interestingly, Newton's language here suggests the strong mental exercise that Descartes calls 'exclusion', as opposed to the weaker one of abstraction. For Descartes, a successful attempt to conceive something while actually separating or excluding another reveals that the two are really distinct, as opposed to being merely conceptually distinct but really identical; see Pr I.62, CSM, 214. Newton's phrase, "supposing God not to exist", suggests the strong mental act of exclusion; he suggests that space may be conceived while actually excluding God, by supposing him not to exist.

[22] *De Gravitatione*, in *Philosophical Writings*, 32.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^c Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

from God, because their capacity to produce sensation cannot be so conceived, and that inconceivability is expressed directly by his definition's third condition.

Interpreting Newton's account: determined quantities of extension and the role of divine action

Yet what exactly are the "determined quantities of extension" endowed with the three conditions that Newton asserts? The question is essential to an understanding of his account of body, but it also has implications for the nature and extent of divine providence, as we will see. It is often supposed that in his creation hypothesis, Newton takes God to create bodies from parts of absolute space itself. For example, Christopher Conn speaks of a body in *De gravitatione* as "nothing more than a divinely-modified region of space".[23] Geoffrey Gorham also takes Newton's determined quantities of extension to be parts of absolute space itself, contrasting the "favored regions of space", which God endows with powers, against the "normal" regions (though on his soft occasionalist interpretation, the favored regions of space are given only powers of producing sensations.)[24] If Newton were seeking some sort of substrate in which properties could inhere, space might initially seem suitable, since as noted earlier, he considers it

[23] Conn, 1999, 316, n. 23. Alan Gabbey allows the possibility without committing to it, in the following passage: "But alternatively, and of equal possibility, the properties of bodies might be the result of God choosing to 'inform' extensions, parts of absolute space, with corporeality and mobility. The parts of absolute space that God can and perhaps does endow with the properties of bodies are as empty of matter as the *materia prima* of the scholastics is void of intelligibility, or bereft of existence. But there is a crucial difference. Each of these parcels of empty extension is a *quid,* and a *quale,* and a *quantum,* whereas *materia prima* is none of these." (Gabbey, "The term *materia* in Newton and the Newtonian Tradition", 16 in proofs). I implied this myself in an earlier article (Kochiras, 2009, 269).

[24] See Gorham, "How Newton Solved the Mind-Body Problem", January, 2011, 22: "Newton proposes that God creates bodies by imposing three conditions on certain regions of space or 'determinate quantities of extension'(DQE).") See also Gorham, "Newton on God's Relation to Space and Time: The Cartesian Framework", September, 2011, esp. 297, where he speaks of "a favored portion of extension".

As a result of taking this line, Gorham understands Newton's account of body as intended to respond to a problem of distinguishing the favored regions of space from the normal ones. The problem (a variant of which was raised by Bennett and Remnant, 1978), may be described by the following two claims. (i) Newton claims that the parts of space are immobile, and therefore the favored portions of space must be distinguishable from the normal parts of space in order to become mobile; yet (ii) the property of impenetrability cannot accomplish the task of making the favored portions of space distinguishable from the normal parts of space, because the normal parts of space are themselves impenetrable to one another precisely because they are immobile. This problem, and the need to resolve it, then motivates Gorham's interpretation of Newton's account of body. In Gorham's view, Newton intends the third condition of his account, i.e., the capacity to produce sensations, to resolve the problem, for in his view, that condition would be superfluous if not intended for that purpose. (Gorham writes, "Condition (3) solves this problem by ensuring that the favored regions of space stand out because God superadds to them something lacking from the unfavored regions: the power to produce sensations." Gorham, January, 2011, 23.)

But the third condition would not be superfluous absent that problem, as I argue in §2.5. Nor is it clear that the problem about distinguishability, which motivates Gorham's account, is genuine. For one thing, if God did modify parts of actual space, surely he himself could distinguish them from one another (as indeed he would have to be able to do, if he were to confer any properties at all upon them.) For another thing, as I argue, Newton's creation story and its associated definition of body does not suppose parts of space itself to be modified. And there is an even more important consideration: even if the problem were genuine, why should we allow the need to resolve it to color our interpretation of Newton's account, given that he himself is not addressing such a problem? Even if the problem were genuine, it should be invoked only to evaluate Newton's account, not to interpret it, since again, Newton himself is not addressing that problem.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

to be more like a substance than an accident. Nevertheless, there are powerful reasons to deny that he supposes God to create bodies by modifying parts of absolute space itself.[25]

The starting point of the creation hypothesis, though hardly decisive, is potentially significant. That starting point is the observation that we can make spaces impenetrable by moving our bodies into them—an action that does not, notably, alter the nature of space itself. Also significant, I think, is the "metaphysical truth" that God "has created bodies in empty space out of nothing"[26]; to square his account with that truth, as he means to do, Newton cannot say that God creates bodies out of space, since space is not nothing. A consideration that should be decisive, however, is the nature of space as he describes it, together with the implications of supposing that actual parts of space figure in his creation story and definition. He described space as being eternal, immutable, immobile, unable to produce effects, and as having parts that are distinguishable but indivisible. To suppose that certain parts of space could be divinely modified, rendered able to produce sensations, solidified and set into motion, is to suppose a full contradiction of Newton's claims . It is to suppose that space is not eternal, because some parts of it may be turned into bodies; that space is not immutable, because some parts could be made impenetrable and able to produce sensations; and that its parts are not immobile and indivisible, because some parts, once made impenetrable, could be torn away from their neighbors and set into motion. And if some parts could be torn away, what exactly would ensue—would space be left with gaps, or would additional space appear to fill the gaps?

These are the sorts of conceptual problems that Newton points to when clarifying the first condition of his definition. Mobility is the first stated condition with which determined quantities of extension are endowed, and since space is immobile, he immediately clarifies that he is not speaking about the parts of space itself, but rather about their quantities: "therefore I did not say that they are numerical parts of space which are absolutely immobile, but only definite quantities which may be transferred from space to space."[27] Significantly, a quantity of some part of space is not identical to the part of space itself—after all, some numerically distinct parts

[25] It should be noted that despite taking parts of space itself to figure in Newton's account of body, Gorham ultimately defends a soft occasionalist interpretation, on which Newton takes the regions of space to be modified only to the extent of temporarily assuming powers to produce sensations in minds. For as noted in §2.5, Gorham argues that the first two conditions of Newton's definition turn out to be superfluous, and the "favored" parts of space, instead of being made actually impenetrable and actually torn away from the "normal" regions of space, are simply "spatial occasions" for God to produce perceptions in minds. Denying that Newton takes the parts of space to be altered and torn apart seems especially important for Gorham since he also argues that space is ultimately identical to God. Therefore, allowing that space could be altered would not only conflict with Newton's claim that space is immutable, it would also imply that God is not immutable; Gorham avoids that implication by arguing that conditions (1) and (2) of the definition "do no independent work".

[26] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 31.

[27] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

of space have the same volume. Thus as Newton's own clarification indicates (a clarification we should keep firmly in mind when he seems to stray from it by employing more abbreviated locutions[28]), it is a mistake to reify his determined quantities of extension, by mistaking them for parts of space itself.[29]

Since Newton associates only quantities with the qualities or powers identified by his three conditions, and not parts of absolute space itself, bodies are constructed from powers alone. Insofar as it is useful to speak in terms of subject and the properties predicated of it, the quantity of any given region of space in which the powers are present may serve as a logical (grammatical) subject, but the utility of such locutions should not lead us to suppose that bodies consist in anything beyond powers. There is nothing like a substrate. Rather, bodies consist in sets of powers, distributed at multiple points of one region of space if the body is resting, or at points of successive regions if the body is moving. This interpretation does require that Newton's first condition, mobility, be considered differently from the other two, in that mobility must apply to something. I therefore suggest that Newton takes bodies (insofar as we can know them) to consist in mobile sets of spatially configured powers for mutual impenetrability and production of sensation. These mobile sets of powers must somehow be unified, so as to maintain their characteristic configurations as they either rest or move through space, and I propose that he assigns the task of unifying them to God. The powers are unified and maintained as enduring configurations by God—by y^e^ divine arm, to borrow a phrase that Newton uses elsewhere.[30] The divine will accomplishes the task that he takes to be performed in the Aristotelian account by prime matter or substrate.

This interpretation fits well with his emphasis upon perceived qualities as the basis of a substance. In one of the explanatory points following his definition of body, he explains that the

[28] At one point, for instance, Newton speaks of the form that God "imparts to space".(*De gravitatione*, in *Newton: Philosophical Writings* 29) Because of such instances, commentators must choose between (i) accepting the surface meaning of such remarks and thus understanding bodies as mobile, solidified regions of space, while paying the price of implying a serious conceptual problem (the question of what would remain, if regions of space could be torn out) as well as conflicts with Newton's own claims (i.e., that space is immutable and immobile, and that his definition concerns definite quantities, not the numerical parts of space); and (ii) avoiding any conflict with his claims that space is immutable and immobile, while paying the price of implying that some of his locutions are abbreviated or careless. I argue for the latter option, as indicated throughout.

[29] My interpretation can be reconciled with the definition that Newton gives of body at the outset of *De gravitatione* (and I thank Eric Schliesser for reminding me, at the conference at Ghent, of the need to reconcile them). As is well known, the bulk of *De gravitatione* consists in a lengthy digression, in which Newton attacks Cartesian physics and addresses various metaphysical questions, including those focused upon here. But Newton begins the manuscript with the intention of treating the weight and equilibrium of fluids and of solids in fluids, and while still engaged in that project, he defines body as "that which fills place" (*De gravitatione*, in *Newton: Philosophical Writings*, 13.) On the interpretation that I develop, that definition can be retained, since a set of spatially distributed powers of mutual impenetrability will repel any other such set; and while such sets do not fill place by actually having parts outside parts, the phenomenal effect is the same.

[30] The phrase is from Newton's second letter to Bentley (17 January, 1692/93; 240 in Turnbull): "Secondly I do not know any power in nature wch could cause this transverse motion without ye divine arm."

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

entities he has described are no less real than bodies and may be called substances bccause "whatever reality we believe to be present in bodies is conferred on account of their phenomena and sensible qualities."[31] And a remark elsewhere in the manuscript, which I discuss in more detail in a subsequent section, points to attributes as the basis of "substantial reality". An interesting implication of my interpretation is that the extension of bodies is parasitic upon the extension of space. Since bodies are extended in virtue of the presence of their constituent qualities or powers in space—a view whose conceptual predecessor is a concept of immaterial spirits as spatially located powers, as noted later[32]—only space is extended in the sense of having parts outside parts, a complete reversal of the Aristotelian view that all extension is corporeal, an attribute of matter.

An objection and response

Still, more needs to be said, because some of Newton's remarks may seem to conflict with the interpretation I have given. In an explanatory remark claiming an advantage for his own account over that of the Aristotelians, he writes, "Extension takes the place of the substantial subject in which the form of the body is conserved by the divine will."[33] This remark, which refers to extension itself, might make one wonder whether Newton does after all mean that God creates bodies by modifying regions of actual space.

I already noted a powerful reason to reject the view that this objection recommends, namely, that it conflicts with Newton's own concept of space and his own clarification that his definition refers to definite quantities, not to numerical parts of space. It should also be acknowledged that the mere mention of extension (space) cannot by itself imply anything, since the mobility condition ensures that absolute space must play some role in Newton's account and hence in any interpretation. Still, the remark figuring in the objection must be explained. To investigate

[31] This claim appears in the second of the four explanatory remarks following Newton's definition of body; *De gravitatione*, in *Newton: Philosophical Writings*, 29.
[32] For a discussion of concepts of spirits and space, see Kochiras, 2012.
[33] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 29. I thank an anonymous referee for urging me to explain how my interpretation can accommodate that remark. The referee also suggests that the following remark may conflict with my claim that the powers comprising bodies are maintained by the divine will: "I do not see why God himself does not directly inform space with bodies, so long as we distinguish between the formal reason of bodies and the act of divine will. For it is contradictory that it [body] should be the act of willing or anything other than the effect which that act produces in space."(Newton, *De gravitatione* in *Newton: Philosophical Writings*, 31.) Newton makes this remark while considering the question of whether God creates bodies directly, as opposed to delegating the task to some intermediary, and he is concerned to distinguish God's action from its effects. The interpretation that I have given does not contravene that distinction. For the powers that God creates, which constitute the body, are the effect of his action and distinct from it; and his action of maintaining those powers in certain configurations is distinct from both the prior action and its effect.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

Newton's meaning, then, I quote the remark in full, along with a second explanatory remark following his definition, which will help illuminate the one particularly at issue.

> That for the existence of these beings it is not necessary that we suppose some unintelligible substance to exist in which as subject there may be an inherent substantial form; extension and an act of divine will are enough. Extension takes the place of the substantial subject in which the form of the body is conserved by the divine will; and that product of the divine will is the form or formal reason of the body denoting every dimension in which the body is to be produced.
>
> Between extension and its impressed form there is almost the same analogy that the Aristotelians posit between prime matter and substantial forms, namely when they say that the same matter is capable of assuming all forms, and borrows the denomination of numerical body from its form. For so I posit that any form may be transferred through any space, and everywhere denote the same body.[34]

In both of these passages, Newton compares his account to the Aristotelian one, but the first repudiates the Aristotelian framework while the second points to a structural similarity between that account and his own.[35] We will need to understand that structural similarity as well as the criticism in order to understand the remark figuring in the objection. Newton's criticism of the Aristotelian account, as elaborated elsewhere in the manuscript, is clear enough: its notions of prime matter or substrate (substantial subject, here) and of a substantial form inhering in that

[34] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 29. These passages are the first and third explanatory remarks following Newton's definition of body. The original of the third explanatory remark (i.e., the second quoted here) reads: "Inter extensionem et ei inditam formam talis fere est Analogia qualem Aristotelici inter materiam primam et formas substantiales ponunt; quatenus nempe dicunt eandem materiam esse omnium formarum capacem, et denomina- tionem numerici corporis a forma mutuari. Sic enim pono quamvis formam per quaelibet spatia transferri posse, et idem corpus ubique denominare." (*Unpublished Scientific Papers of Isaac Newton,* 107.)

[35] An interesting interpretation of *De gravitatione* has been given by Benjamin Hill, who does not see the mere structural similarity that I take Newton to assert between his view and the scholastic one, but rather sees significant scholastic content in Newton's ideas ("Newton's *De Gravitatione et Aequipondio Fluidorum* and Lockean Four-Dimensionalism", 2003.) One point of agreement between my view and Hill's is that both deny that the determined quantities of extension figuring Newton's account of body are regions of actual space. Apart from that, however, our views differ in a number of ways. For one thing, Hill understands the account in terms of *extensio* interpreted as potentiality. He argues that Newton retains "the metaphysical structures of the Scholastics' hylomorphism but substituted into those structures extension for prime matter and impenetrability + mobility for substantial form."(Hill, 2003, 317) On Hill's analysis, these substitutions are possible because Newton's *extensio* (which is a quantity, and thus distinct from space itself) is similar to the Scholastics' prime matter in a crucial way: "In Newton's thought, extension was, like prime matter, *pura potentia*". (Hill, 2003, 318; see also 321: "Although he did not strictly adhere to it...Newton seems to have distinguished *extensio* from *spatium*. *Spatium* denoted physical space whereas *extensio* denoted the abstract and metaphysical extensive quantity.")

Although his interpretation is ingenious, I am not convinced by it, and the difficulties I see are instances of an objection he anticipates and addresses, namely, that he has exaggerated Newton's scholasticism (see Hill, 320-321). Specifically, I am not convinced that Newton distinguishes *extensio* and *spatium*, as Hill claims, or that he understands the former as *pura potentia*. In connection with this, Hill's interpretation does not easily accommodate Newton's claim that the scholastic notion of prime matter is unintelligible. If we suppose that Newton understood prime matter as *pura potentia*, it is not clear why he would attack it as unintelligible (particularly if we also suppose that Newton understood the determined quantities of extension figuring in his own account of body as *potentia*). His charge that prime matter is an unintelligible notion is explained, however, if we suppose that he understands and represents it uncharitably (as he often represents Descartes) as a propertyless substrate that is an actual, component in substances; and his attack upon the scholastic account suggests that that is the way he understands it, as I indicate in §4. For instance, Newton writes, "Further, they attribute no less reality in concept (though less in words) to this corporeal substance regarded as being without qualities and forms, than they do to the substance of God; abstracted from his attributes." (Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 32.) Here Newton takes the Scholastics to explain bodies in terms of a propertyless, corporeal substrate, and he criticizes them for attributing reality to this concept.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

prime matter are unintelligible.[36] This charge motivates the advantage he claims for his own account: since extension "takes the place of the substantial subject", he avoids the unintelligible notion of prime matter.

Turning to the structural similarity, Newton takes extension (space) in his own account to be analogous to prime matter in the Aristotelian account; and form in his account (which he also refers to as the product of the divine will) to be analogous to their substantial form. Before proceeding, we must ask what could he mean by 'form' in connection with his own account. I think he means 'form' to refer to the extent and shape of the configured set of powers. For in a limited class of cases, the Aristotelians do take form to be little more than shape, and that is a use of the term that Newton can accept, even as he rejects the notion of substantial form more generally. Thus, when he writes that the form of the body is conserved by the divine will, he means that the spatial configuration of the set of powers is maintained by God's action, as I argued earlier.[37]

Proceeding, then, we next need to understand the relation Newton sees between prime matter and substantial form in the Aristotelian account, since that will enable us to understand the relation he asserts between extension and form in his own account.[38] He represents the Aristotelians to be saying the following. Since prime matter can be associated with any form, its association with any body, via a particular form, is merely contingent; and so it is the substantial form that individuates the body.[39] That is to say, although prime matter facilitates a body's

[36] See, for instance 31-32 of *De gravitatione*. It may be remarked that the unintelligibility of Aristotelian substratum is due at least in part to Newton's portrayal of it as something already complete in itself, as opposed to an incomplete material principle, which together with a substantial form contributes to the production of a complete, accident-bearing substance. Also, Newton's representation of prime matter as lacking all qualities overlooks the view, held by all Scholastics other than strict Thomists, that prime matter possesses the capacity for extension (*extensio in potentia*), a point I owe to Dennis Des Chene. And the Scholastics did grapple with the question about prime matter's intelligibility. It may also be remarked that although Newton sometimes uses the term 'inhere' (or its cognates) in his own assertions—notably in Definition 3 of the *Principia*, which defines the *vis insita* (inherent force), also called the *vis inertia* (force of inertia)—he is not there employing the scholastic sense of the term. For as is eventually made clear via the explanatory remarks at the end of Rule 3, Newton means to contrast the *vis insita/ vis inertiae* against relational forces, notably the gravitational force. Unlike gravity, the *vis insita/ vis inertiae* is monadic—it belongs to the body itself.

[37] This reading is supported by his remark, at the end of the first passage, that the form denotes each dimension in which it is produced. That is to say, the form or spatially configured set it marks out the same dimension (quantity of space), as it moves through numerically distinct parts of space. To borrow *Principia* terminology, the set of powers provides a sensible measure of each space it occupies, by reflecting other such sets, including light.

[38] Alan Gabbey (forthcoming, 10), commenting upon both this passage and a similar remark that Newton makes in a much later text (Add 3965 (no. 13), ff. 422r) writes, "Right to end of his life Newton saw an analogy between the Peripatetic couple, *materia prima* and *forma substantialis,* and the Newtonian couple, the endlessly transmutable matter common to all bodies and their properties, phenomena available to one or other of the senses."(Gabbey, "The Term *Materia* in Newton and in the Newtonian Tradition", forthcoming, 12). I do not mean to imply that Gabbey accepts my interpretation of Newton's account of body, but I find his remark illuminating.

[39] Since matter can assume all forms, Newton implies, then if matter rather than form individuated substances, there would be only a single substance persisting, no matter how dramatic the change in qualities. As a point of clarification that I owe to Dennis Des Chene, Newton incorrectly implies in this passage that there was agreement among the Scholastics about the

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

existence (since both prime matter and substantial form are needed for the body to exist), it never really belongs to the body because its association with that body is contingent; and therefore, to refer to the body is really to refer to its form.

Newton sees the same sort of relation in his own account, writing that "any form may be transferred through any space, and everywhere denote the same body." Space facilitates a body's existence, in that the body's powers must be distributed in space—for as noted earlier, no being can exist without being somehow related to space. Yet any given region of space may be associated with any body, since any body may occupy or pass through it; and since that region's association with the body (set of powers) is contingent, it cannot be said to belong to the body. This is Newton's point when he writes that the form denotes the same body, even as it is transferred through different spaces. Thus the interpretation that I have given can make sense of the passages discussed. (And if fact it makes better sense of them than does the interpretation claiming bodies to be divinely modified parts of actual space. That interpretation cannot account for the contingent, transitory relation the passages assert to hold between a part of space and the form, for if a part of space were modified so as to become a body, its relation to the form would not be contingent or transitory.)

The account of body and the extent of God's providence

In another of the explanatory remarks following the definition of body, Newton states that the entities he has described subsist "through God alone".[40] The interpretation I have given provides a specific way of understanding this: the entities subsist through God alone in that the sets of powers are unified and maintained in their configurations by divine action. Since this action is direct, God's providence is much greater than if he merely concurred with the bodies' continued existence. Still, Newton also leaves ample room for secondary causation, for as indicated earlier, he sees the account of body and thus God's direct action as limited to corpuscles. This suggests a view similar to that found in a much later text, Query 31 of the *Opticks*. Query 31 sidesteps the problem of cohesion at the sub-corpuscular level by suggesting that corpuscles are created by God, but it speculatively attributes the cohesion of aggregate

principle of individuation. Des Chene further explains (in correspondence) that there was some agreement among them that "substantial form *would* individuate corporeal substance, were it not that matter can exist, by the absolute power of God, without form and even without quantity".

[40] *Ibid.* Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 29.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^e^ Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

bodies to interparticulate forces, and thus to secondary causes.[41] Here too, by restricting his account of bodies to corpuscles, Newton leaves the cohesion of aggregate bodies to secondary causes.

The role that Newton assigns to God in *De gravitatione* therefore falls considerably short of occasionalism. This is consistent with the expectations that he evinces in other texts. In a letter of 1680, Newton writes, "Where natural causes are at hand God uses them as instruments in his works".[42] And as I have argued elsewhere, Newton never endorses the hypothesis that God causes gravitational effects directly, and his ongoing search for an explanation expresses his expectation of secondary causes.[43]

I therefore disagree with the interpretation defended recently by Gorham, who attributes occasionalism to Newton, albeit a soft sort.[44] The occasionalism is soft in that God does not cause perceptions in minds directly, instead endowing varying regions of space with the power to do so, in a continuous creation of matter.[45] Yet it is still a kind of occasionalism, because Gorham argues that that the first and second conditions of Newton's definition of body are superfluous, doing "no independent work of their own",[46] and that bodies consist in only the powers to produce sensations. Regions of space are the "spatial occasions" for the sensations, and God creates matter continuously by creating the powers to produce sensations in varying regions of space.[47] Gorham claims a powerful advantage for his interpretation: it implies that Newton solves the mind-body problem, avoiding problems about mental causation "by embracing a quasi-idealistic ontology of matter."[48] Yet his interpretation requires us not only to accept that conditions (1) and (2) of Newton's definition are superfluous, but also that condition (3), the power to produce perceptions in minds, is not merely necessary for body-hood but also sufficient. Gorham reaches this latter conclusion partly through his reading of the comment that Newton adds to this third condition—that it is not surprising that bodies have the power to cause

[41] An illuminating discussion of Locke and the foundational problem about cohesion may be found in James Hill (2004), "Locke's Account of Cohesion and its Philosophical Significance".
[42] Newton to Burnet, 1680; Newton, *The Correspondence*, II, 334.
[43] See Kochiras, 2009, 2011.
[44] Gorham indicates that he sees Newton as belonging to a tradition that locates the ground of causation in God's will (Gorham, January, 2011, 25).
[45] See Gorham, January, 2011, 24.
[46] Gorham, January, 2011, 24.
[47] See Gorham, January, 2011: "The continuous creation of matter amounts simply to the distribution within space of God's power to produce sensations"(24); and "various quantities of extension are the mere 'spatial occasions' for God to bring out our perceptions in the successive and law-like ways we associate with moving bodies."(25).
[48] Gorham, January 2011, 30.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

perceptions in minds, "since the description of their origin is founded on this".[49] Yet there is a natural reading of that remark which does not require either dismissing the definition's first two conditions as superfluous or supposing the third to be sufficient. That natural reading, which I explained earlier, is simply that if bodies lacked the power to produce sensations, we could never have any ideas of them. The remark is an instance of Newton's oft-repeated acknowledgement that we can know only perceived qualities, not the "essential and metaphysical constitution" of things.[50] Since I reject the occasionalist interpretation, I also reject Gorham's conclusion that "Newtonian bodies do not seem to qualify as self-standing substances".[51] On my interpretation, Newton considers bodies to be created substances. This is a desirable result, since bodies would have to be substances in order for Newton to accept a substantial distinction between mind and body—and he does, as I argue elsewhere.

In closing, I suggest that the account of body Newton develops in *De gravitatione* might have indirectly helped facilitate a concept belonging to his later rational mechanics, that of point mass. On the interpretation I have given, his concept of body has as its conceptual ancestor a spirit which consists in causal powers, which lacks parts outside parts, and which is extended only in the derivative sense that its constituent causal powers are present in some extension. An entity consisting in spatially present causal powers, as opposed to one possessing parts outside parts, may more easily be conceived as existing in a larger or smaller area—even as contracted to a point. Thus the bodies of *De gravitatione*, which consist in powers of mutual impenetrability or resistance, might have helped facilitate Newton's realization that mass can be considered at a point. Or at least, because they lack parts outside parts, such bodies would not stand in the way of that realization.

---

[49] *De gravitatione*, 29. There is another passage that Gorham interprets as showing that Newton takes condition (3) to be sufficient as well as necessary for being a body. In that passage, Newton is attacking the Cartesian view of matter:

> "Let us abstract from body (as he demands) gravity, hardness, and all sensible qualities, so that nothing remains except what pertains to its essence. Will extension alone then remain? By no means. For we may also reject that faculty or power by which they [the qualities] stimulate the perceptions of thinking things. For since there is so great a distinction between the ideas of thought and of extension that it is not obvious that there is any basis of connection or relation [between them], except that which is caused by divine power, *the above capacity of bodies can be rejected while preserving extension, but not while preserving their corporeal nature*."(Newton, *De gravitatione* in *Isaac Newton: Philosophical Writings*, 33-34; emphasis added)

Commenting upon this passage, and quoting the italicized portion, Gorham writes, "So, the capacity to produce sensations in minds is sufficient and necessary for a quantity of space to possess the nature of body. This explains why Newton privileges condition (3) when he introduces his theory of creation: "The description of their [bodies'] origin is founded on this" (*De Grav* 29)."(Gorham, January, 2011, 24.) I do not see how Newton's remarks imply that condition (3) is sufficient as well as necessary for body-hood, as Gorham takes it to do. There is certainly a way of understanding the passages that does not imply any such thing. As I read the passage, Newton is saying that if one mentally abstracts qualities such as hardness away from a body, one has abstracted away only something that is necessary to body, not everything, since bodies also have the power to produce sensations. He is saying that condition (3) is necessary to body-hood, but he is not saying that it is sufficient.

[50] *De gravitatione*, 27; Cf. *Principia*, 942.

[51] Gorham, January 2011, 24.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

References

Bennett, J. and Remnant P. (1978). "How Matter Might First be Made". New Essays on Rationalism and Empiricism. Ed. C. E. Jarrett, J. King- Farlow, and F. J. Pelletier. *Canadian Journal of Philosophy*, Supplementary Volume 4: 1-11.

Cohen, I.B. 'Versions of Isaac Newton's first published paper', *Archives Internationales d'Histoire des Sciences*, 11, 1958: 357-75.

Cohen, I. B. (1999). A guide to Newton's Principia (with contributions by M. Nauenberg, & G. Smith). In I. Newton, *The Principia: Mathematical principles of natural philosophy* (I. B. Cohen, & A. Whitman, Trans.) (pp. 1–370). Berkeley: University of California Press.

Conn, Christopher H. "Two Arguments for Lockean Four-Dimensionalism". *British Journal for the History of Philosophy*, ISSN 0960-8788, 10/1999, Volume 7, Issue 3, pp. 429 – 446.

Dempsey, Liam, "Written in the flesh: Isaac Newton on the mind–body relation", *Studies in History and Philosophy of Science*, Vol. 37, No.3 (2006), pp. 420-441.

Descartes, R. (1985). *The philosophical writings of Descartes*, Vols. 1 & 2 (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.). Cambridge: Cambridge University Press. (Abbreviated CSM).

Gabbey, Alan. "The Term *Materia* in Newton and in the Newtonian Tradition" (forthcoming; pagination here refers to proofs and may differ upon publication).

Garber, Daniel (2001). *Descartes Embodied*, Cambridge: Cambridge University Press.

Gorham, G. (2011a) "How Newton Solved the Mind-Body Problem". *History of Philosophy Quarterly* 28: 21-44.

Gorham, G. (2011b) "Newton on God's Relation to Space and Time: The Cartesian Framework". *Archiv für Geschichte der Philosophie*, 93: 281-320.

Grant, Edward. *Much Ado about Nothing: Theories of Space and Vacuum from the Middle Ages to the Scientific Revolution*. New York: Cambridge University Press, 1981.

Henry, John: "A Cambridge Platonist's Materialism: Henry More and the Concept of Soul", *Journal of the Warburg and Courtauld Institutes*, Vol. 49 (1986), pp. 172-195.

Hill, B. (2003). Newton's De Gravitatione et Aequipondio Fluidorum and Lockean Four-Dimensionalism". *British Journal for the History of Philosophy* 11: 309-321.

Hill, J., 2004, "Locke's Account of Cohesion and its Philosophical Significance," *British Journal for the History of Philosophy*, 12 (4): 611 – 630.

This paper, presented on June 14, 2012 at the 7th Quadrennial Fellows Conference, is an excerpt from
Kochiras: "By y^{e} Divine Arm: God and Substance in *De gravitatione*", *currently under review.*

Hooke, Robert: Lectures *Potentiae Restitutiva*, or of Spring, Explaining the Power of Springing Bodies. Printed for John Martyn, Ptrinter to the Royal Society, at the Bell in St. *Pauls* Church-Yard, 1678.

Kochiras, H. (2009). Gravity and Newton's Substance Counting Problem, *Studies in History and Philosophy of Science*, 40(3): 267-280.

Kochiras, H. (2011). "Gravity's Cause and Substance Counting: Contextualizing the Problems", *Studies in History and Philosophy of Science,* 42(1): 167-184

Kochiras, H. "Spiritual Presence and Dimensional Space beyond the Cosmos", *Intellectual History Review*, 22(1) 2012: 41–68.

Kochiras, H. "Newton and the Doctrine of Holenmerism", n.d. (manuscript)

Locke, John, *An Essay Concerning Human Understanding*, edited Peter H. Nidditch, New York: Oxford
University Press, 1975.

McGuire, J. E. (1978). "Existence, Actuality and Necessity: Newton on Space and time". *Annals of Science*, 35, 463-508.

McGuire, J.E. and Slowik, Edward, "Newton's Ontology of Omnipresence and Infinite Space", *Oxford Studies in Early Modern Philosophy* (forthcoming).

More, Henry. *Philosophical Writings of Henry More*, ed. F.I. MacKinnon. New York: Oxford University Press, 1925.

Newton, I. (2004). Newton: Philosophical writings (A. Janiak, Ed.). Cambridge: Cambridge University Press.

Newton, I. (1962). *Unpublished scientific papers of Isaac Newton* (A. R. Hall, & M. B. Hall, Eds. & Trans.). Cambridge: Cambridge University Press.

Newton, I. (1959–1971). *Correspondence of Isaac Newton* (H. W. Turnbull, J. F. Scott, A. R. Hall, & L. Tilling, Eds.) (7 vols.). Cambridge: Cambridge University Press.

Newton, I. (1952). *Opticks, or A treatise of the reflections, refractions, inflections and colors of light.* Based on the fourth edition of 1730. New York: Dover.

Nolan, Lawrence, "Reductionism and Nominalism in Descartes's Theory of Attributes". *Topoi* 16: 129–140, 1997.

Reid, Jasper, "The Spatial Presence of Spirits among the Cartesians" *Journal of the History of Philosophy*, Volume 46, Number 1, January 2008, pp. 91-117.

Slowik, Edward. "Newton's Neo-Platonic Ontology of Space", *Foundations of Science* (forthcoming).

Slowik, Edward. Newton's Metaphysics of Space: A "Tertium Quid" betwixt Substantivalism and Relationism, or Merely a "God of the (Rational Mechanical) Gaps"? *Perspectives on Science*, Volume 17, Number 4, Winter 2009, pp. 429-456.

Stein, Howard, "Newton's Metaphysics", in I. Bernard Cohen and George Smith, editors, *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press, 2002.

GÖDEL ON TRUTH AND PROOF:
Epistemological Proof of Gödel's Conception of the Realistic Nature of Mathematical Theories and the Impossibility of Proving Their Incompleteness Formally

Dan Nesher, Department of Philosophy University of Haifa, Israel

> No calculus can decide a philosophical problem. A calculus cannot give us information about the foundations of mathematics. (Wittgenstein, 1933-34: 296)

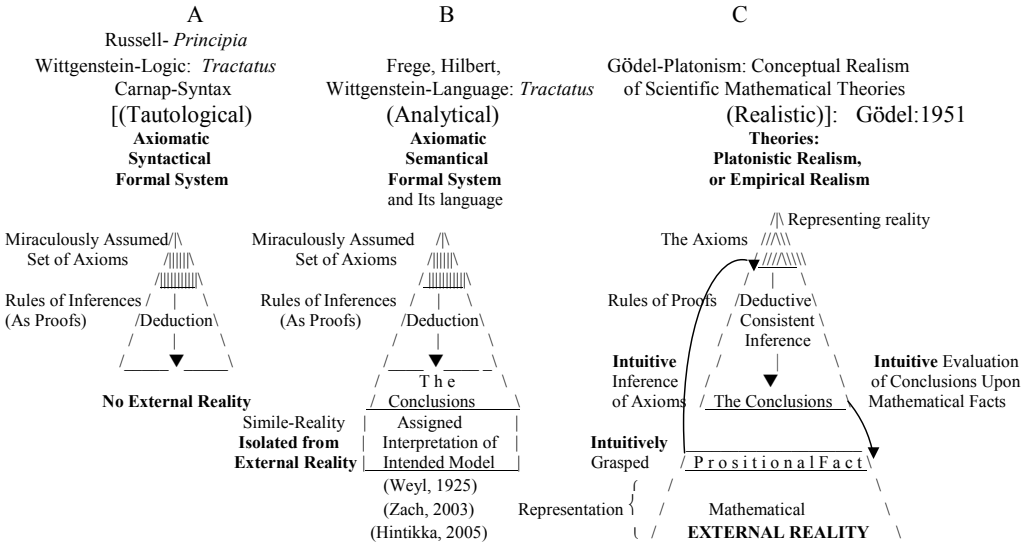
1. Introduction: Pragmaticist Epistemological Proof of Gödel's Insight of the Realistic Nature of Mathematical Theories and the Impossibility of Proving Their Incompleteness Formally

In this article, I attempt a pragmaticist epistemological proof of Gödel's conception of the realistic nature of mathematical theories representing facts of their external reality. Gödel generated a realistic revolution in the foundations of mathematics by attempting to prove formally the distinction between complete formal systems and incomplete mathematical theories. According to Gödel's Platonism, mathematical reality consists of eternal true ideal facts that we can grasp with our mathematical intuition, an analogue of our sensual perception of physical facts. Moreover, mathematical facts force us to accept intuitively mathematical true axioms, which are analogues of physical laws of nature, and through such intuition we evaluate the inferred theorems upon newly grasped mathematical facts. However, grasping ideal abstractions by means of such mysterious pure intuitions is beyond human cognitive capacity. Employing pragmaticist epistemology, I will show that formal systems are only *radical abstractions* of human cognitive operations and therefore cannot explain how we represent external reality. Moreover, in formal systems we cannot prove the truth of their axioms but only assume it dogmatically, and their inferred theorems are logically isolated from external reality. Therefore, if Gödel's incompleteness of mathematical theories holds, then we cannot know the truth of the basic mathematical facts of reality by means of any formal proofs. Hence Gödel's formal proof of the incompleteness of mathematics cannot hold since the truth of basic facts of mathematical reality cannot be proved formally and thus his unprovable theorem cannot be true. However, Gödel separates the *truth* of mathematical facts from mathematical *proof* by assuming that mathematical facts are eternally true and thus, the unprovable theorem seems to be true. Pragmatistically, realistic theories represent external reality, not by formal logic and not the abstract reality, but by the *epistemic logic* of the complete proof of our perceptual propositions of facts and realistic theories. Accordingly, it can be explained how all our knowledge starts from our perceptual confrontation with reality without assuming any *a priori* or "given" knowledge. Hence, mathematics is also an empirical science; however, its represented reality is neither that of *ideal objects* nor that of *physical objects* but our operations of counting and measuring physical objects which we perceptually quasi-prove true as mathematical basic facts (Nesher, 2002: V, X).

2. Gödel's Platonism and the Conception of Mathematical Reality with Its True Conceptual Facts

Gödel's basic insight of the realistic nature of mathematics that it is a science represents mathematical reality and not just a conventional formal system. Yet, Gödel's Platonist mathematics is an abstract science representing ideal true mathematical reality though analogical to the empirical sciences (Gödel, 1944). As a *metaphysical realist*, Gödel separates the mathematical reality of abstract true facts from formal proofs, and it is only by pure intuition that we can grasp these facts. Figure 1 presents a schema of Gödel's different conceptions of logic and mathematics:

[1] The Gödelian Epistemology of Three Conceptions of Logic and Mathematics:



Gödel's tri-partitions are between (A) *Complete* Analytic Formal Systems with their formal syntactic *tautologies*, (B) *Complete* Formal Semantic *analyses*, and (C) the *Incomplete* Realistic Theories of *conceptual* mathematics (Gödel, 1951: 319-323; Poincaré, 1902: Chap. I).

> The two significations of the term *analytic* might perhaps be distinguished as tautological and analytic (Gödel, 1944:139, n. 46).

Epistemologically the *tautological* and *analytic* of *complete formal systems* are, respectively, ***syntactically* closed** upon their fixed axioms and formal rules of inference and ***semantically closed*** upon axioms, formal rules, and the assigned model. The *realist incomplete theory* is only ***relatively closed*** upon its relative proof-conditions, the formal proofs, the operations of pure intuition, and conceptual facts of external reality (Nesher, 2002: X). Since Gödel's mathematical theories are regarded as axiomatic formal systems with formal inferences, yet their external reality can be grasped only by pure intuition (Gödel,

1931a: 203, 1964: 268).

For Gödel, pure mathematical intuition has three functions: (1) to grasp the true ideal mathematical facts of mathematical reality, (2) to enforce by these ideal facts to accept the true axioms of mathematical theories in order to infer the theorems formally, and (3) to evaluate how the theorems represent truly facts of mathematical reality (Gödel, 1953-54?: fn. 34; Nesher, 2001a, 2010). Gödel's conception of mathematical intuition is based on his mathematical experience, which he calls the "psychological fact of the existence of an intuition," but as a "given" without any explanation.

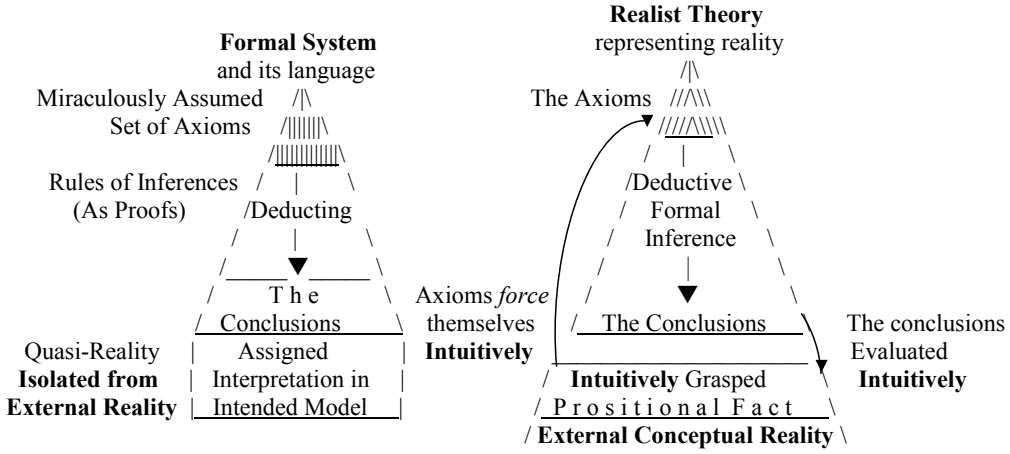
> However, the question of the objects of mathematical intuition (which incidentally is a replica of the question of the objective existence of the outer world) is not decisive for the problem under discussion here. The mere psychological fact of the existence of an intuition which is sufficiently clear to produce the axioms of set theory and an open series of extensions of them suffices to give meaning to the question of the truth or falsity of propositions like Cantor's continuum hypothesis (Gödel, 1964: 268).

How with mathematical intuition we grasp pure meanings of mathematical propositions is the essential problem to the possibility of Gödel's conceptual realism (Gödel, 164:268).

3. Gödel's Incomplete Distinction between Formal Systems and Realistic Theories

Gödel revolutionized the conception of the nature of mathematics through his distinction between complete logical formal systems and incomplete mathematical theories (Gödel, 1931:195, 1964). However, he did not conclude this revolution, because of his acceptance the formalist methods of mathematical proofs and the subjective conception of pure intuition owing to his Platonist realism that motivated this revolution (Gödel, 1931:#1).

[2] **Epistemological Gap between Logical Formal Systems and Mathematical Theories**



The difference between formal system and realist theory lies in their proof-conditions when the formal system is by definition **hermetically closed** upon its fixed formal proof-conditions without relation to external reality; the mathematical realistic theory is **relatively closed** upon its proof-conditions: the mathematical facts of external reality, the formal inferences, and the pure intuitions that complete the representation of reality, while the axioms change by our continually grasping new mathematical facts. Yet, the formal systems are *artificially* abstracted from human mathematical operations and cannot explain them, and thus they can never be "ideal machines" by lacking any human cognitive self-consciousness and self-controlled operations upon reality (Gödel, 1931: 195 & n. 70; 1951: 310; Feferman, 2006; Putnam, 2011; Penrose, 2011). Apparently Gödel did not completely conceive his epistemological revolution of the realistic nature of mathematics and considered the three classes of logico-mathematics**,** A, B, and C, as formal systems, while neglecting the essential distinction between formal systems and mathematical theories.

> The development of mathematics toward greater precision has led, as is well known, to the formalization of large tracts of it, so that one can prove any theorem using nothing but a few mechanical rules. The most comprehensive formal systems that have been set up hitherto are the system of *Principia mathematica* (*PM*) on the one hand and the Zermelo-Frankel axiom system of set theory . . . on the other. These two systems are so comprehensive that in them all methods of proof today used in mathematics are formalized, that is, reduced to a few axioms and rules of inference. (Gödel, 1931: #1; cf. 1931a; Kleene, 1967: 253)

Gödel's incompleteness theorem essentially shows that *PM* and ZF are mathematical theories, not formal systems; however, since they use formal inferences, then without the help of mathematicians' conceptual intuition, those systems are isolated from mathematical reality. According to pragmaticist epistemology, the formal inference is only one component of the epistemic logic which includes also the Abductive and Inductive material inferences of the complete proof enable also to prove the basic mathematical facts of external reality. Yet, even after proving the incompleteness of mathematics, Gödel still oscillated between mathematics as axiomatic formal systems and as scientific theories, and thus he could not complete his realistic revolution of mathematics (Gödel, 1953-54? II; Feferman, 1984: 9-11).

4. Gödel's Paradoxical Formal Proof of Incompleteness, Based on Separating Truth from Its Proof

If Gödel's incompleteness holds, then mathematics is theory and a not formal system so, can Gödel prove formally his incompleteness in mathematical theory that cannot prove formally true theorems (Hintikka, 2000: V)? Gödel's formal proof of incompleteness is actually an "arithmetization of syntax," which attempts to prove his epistemological conception of the nature of mathematics. But Gödel's incompleteness is a general claim that can be proved only epistemologically, and not through any specific

theory about itself. It could be that in respect of a special mathematical theorem it can be prove that a specific theory (e.g., PM or ZF) is incomplete in respect to specific propositions and the given true mathematical facts; but it cannot provide a general proof of the nature of mathematics (Gödel, 1944:121).

Gödel arithmeticized the proof of the undecided proposition G1: "I am unprovable," by means of a *metamathematical description* in order to prove this unprovable mathematical proposition, "We therefore have before us a proposition that says about itself that it is not provable [in *PM*]" (Gödel, 1931: 151). The question is whether this formal proof can be considered proof of G1: "I am unprovable"? There are two problems here: (1) Can at all there be metalanguages, since meta-descriptions of mathematical languages can , at most, describe physical-syntactical signs, following Tarski, and not their meaning-contents, which we can only interpret, yet not in abstract models but in respect to experience (Wittgenstein, 1921, 1933-34: II.12; Gödel, 1953-54?: fn.34, p.203: Nesher, 1987, 2002: V)? (2) Can G1 be meaningful and "contentually true" that eventually represents a mathematical true fact (Gödel, 1931a: 203)?

If G1: "I am unprovable" is proved formally *true* in *PM*, then its claim of being unprovable is *false* because it was proved true [in *PM*] and cannot be unprovable, but when G1 is *false* then being unprovable in *PM* is *true* as it claims, and thus presenting a paradox like the liar paradox, and Gödel's trick of using a kind of paradoxical argument fails.

> The analogy of this argument with the Richard antinomy leaps to the eye. It is closely related to the "Liar" too.[14] (Any epistemological antinomy could be used for a similar proof of the existence of undecidable propositions). (Gödel, 1931: 149)

Since any epistemological antinomy is void of truth, this means that its proof is also void of truth. It seems that Gödel felt this difficulty, and his way out of this paradoxical situation is to locate the *proof* at the metamathematical arithmetical language and thus separate this formal proof from the language of G1 with the assuming truth of its bizarre meaning.

> From the remark that [R(q);q] says about itself that it is not provable, it follow at once that [R(q);q] is true, for [R(q);q] is indeed unprovable (being undecidable). Thus, the proposition that is undecidable *in the system PM* still was decided by metamathematical considerations. (Gödel, 1931: 151)

Why did Gödel take recourse in this "epistemological antinomy" as a trick and not proving the incompleteness of *PM* by showing that propositions "of the type of Goldbach or Fermat" are unprovable in it (Gödel1931a: 203)? It seems that Gödel intended a general proof of the nature of all mathematical theories in respect of their infinite mathematical reality (Agazzi, 1974: 24). Gödel's Platonist realism leads him to formulate his proof with the suffix *able* as his "provable" and "unprovable" terms. This means that

since there are eternal and infinite true mathematical facts that eventually can be grasped by pure intuition, they are either provable or unprovable in any mathematical theory (Hintikka, 2000:29). In such Platonic epistemology, *truth* in reality and *proof* in theories are *separated*, which enables Gödel to separate the *proof* of G1 from the *truth* of the mathematical fact it is to represent, in order to avoid the paradox in proving his incomplete theorem of being "closely related to the 'Liar.'"

> Finally it should be noted that the heuristic principle of my construction of undecidable number theoretical propositions in the formal systems of mathematics is the highly transfinite concept of 'objective mathematical truth' as *opposed* to that of 'demonstrability' . . . , with which it was generally confused before my own and Tarski's work (Gödel in a letter to Wang, Dec. 7, 1967, in Wang, 1974: 9; Feferman, 1984: 106-107; Franzèn, 2005: 2.4).

Hence, Gödel leans on the distinction between the liar proposition P^L: "I am lying" and the unprovable proposition P^U: "I am unprovable" since in the former we reach the liar paradox that if it is true then it is false and vice versa, whereas there is no such paradox of truth and falsity in the latter, since proof and truth are separated (Gödel, 1934 #7, 1951: 322-323; Hintikka, 2000:35-36; Devlin, 2002).

> So we can see that the class α of numbers of true formulas cannot be expressed by the propositional function of our system, whereas the class β of provable formulas can. Hence $\alpha \neq \beta$ and if we assume $\beta \subseteq \alpha$ (i.e., every provable formula is true) we have $\beta \subset \alpha$, i.e., there is a proposition A which is true but not provable. ~A then is not true and therefore not provable either, i.e., A is undecidable (Gödel, 1934: 363).

Generally, Gödel separates the truth of mathematical facts, which can be grasped intuitively, from the formal *proof* of propositions in mathematical theories and thus also, he can separate the attempted formal *proof* of G1 from its seemingly representing the *truth* of a fact in the mathematical reality of *PM*. Leaning on his Platonistic realism he could do it in order to avoid the possibility that G1 would be both true and false like the Tarskian liar proposition.

> Thus if truth for number theory *were* definable within itself, one could find a precise version of the liar statement, giving a contradiction. It follows that truth is not so definable. But provability in the system *is* definable, so the notions of provability and truth must be distinct. In particular if all provable sentences are true, there must be true non-provable sentences. The self-referential construction applied to provability (which is definable) instead of truth, then leads to a specific example of an undecidable sentence (Feferman, 1984: 106).

However, if the notions of truth and proof are not separated there are no "true non-provable sentences" and "the self-referential construction" of G1 leads to an "epistemological antinomy," a kind of the liar paradox. Metaphysical realists, such as Platonists and formal semanticists (e.g., Tarski), assume that truth is independent of proof and, by the bivalence of truth values, the principle of excluded middle, identify truth with reality, yet, not for complete formal systems (Gödel, 1929: 63; Penrose, 2011: 342-343).

Pragmaticists, however, show that for humans the truth and falsity of propositions consist only of that which we have already proved as such, since we cannot know truth from a Godly perspective (Nesher, 2002: V). Since there is no separation between truth and being proved, then we have to drop the expressions "provable" and "unprovable" from our epistemology. This terminology belongs to Metaphysical Realism, such as Gödel's Conceptual Realism, Popper's absolute truth, among others, in distinction from Pragmaticist Representational Realism (Nesher, 2002: III, V, VIII).

Therefore, without being proved true or false, propositions remain doubtful, and since no one has proved the truth or the falsity of the *liar* proposition, it is doubtful and there cannot be any paradox (Nesher, 2002: V). Hence the separation of truth from proof is epistemologically untenable and so also the separation between the liar paradox and the unprovable-provable antinomy, and thus, with the doubtful *unprovable* proposition we cannot prove anything (Hintikka, 2000:31-35). Although Wittgenstein sensed the paradoxical difficulty in Gödel's alleged proof of incompleteness, he could not explain it without having an epistemology of truth (Wittgenstein, 1937; Nesher, 1992; Floyd and Putnam, 2000; Floyd, 2001; Berto, 2009: # 9).

> How can Gödel prove that his crucial proposition is not logically provable by using the very same logic? And how we can know that the proposition in question is true if we cannot prove it? (Hintikka, 2000:29)

What, then, is the meaning of G1 if it were proved to represent a conceptual true fact in mathematical reality? And can we specify this true fact that the alleged meaning-content of G1 represents? Indeed, there is no mathematical fact that G1 represents, since it is not a proposition with real subject matter and clear content and if anything at all, it has only a shadowing meaning (Gödel, 1931a: 203; Weyl, 1949: 51; Feferman, 1984: 106). However, if G1: "G1 is unprovable" is void of real meaning and thus cannot be "contentually true" then it cannot represent any intended "mathematical objects or facts exist," according to Gödel's criticism of the syntactic conception of mathematics (Gödel, 1931a: 203, Gödel, 1953-54?: #30; Agazzi, 1974: 24; Feferman, 1984: 103). Hence the arithmeticized proof of G1 is only mechanically connected to the object language and has nothing to do with its meaning (Tarski, 1944; Nesher, 1987, 2002: V; Floyd, 2001: III). Then if G1 can be proved formally, any sentence can be proved emptily and the system or theory in which it is proved is inconsistent (Gödel 1931a: 203).

> This formulation of the non-feasibility of the syntactic program (which also applies to finitary mathematics) is particularly well suited for elucidating the question as to whether mathematics is void of content [in the sense that no mathematical objects or facts exist]. For, if *prima facie* content of mathematics were only a wrong appearance, it would have to be possible to build up mathematics

> satisfactorily without making use of this "pseudo" content. (Gödel, 1953-54?: #30; Hintikka, 2000: 29)

However, the meaning-contents of scientific theories are based on our experiential confrontation with external reality and mathematical reality, as well. Thus, the basic facts of mathematical reality cannot be proved formally in theory from its axioms and the question is how we prove their truths and whether we can grasp their truths by pure mathematical intuition (Gödel, 1944: 21).

> It is turned out that (under the assumption that modern mathematics is consistent) the solution of certain arithmetical problems requires the use of assumptions essentially transcending arithmetic, i.e., the domain of the kind of elementary indisputable evidence that may be most fittingly compared with sense perception. (Gödel, 1944: 121; cf. Gödel, 1953: #34)

This Gödel insight fits the pragmaticist understanding of the role of epistemic logic proofs in all empirical sciences, mathematics included (Gödel, 1947: 182-183, 1964: 268-269; Nesher, 2002, 2007; Chihara, 1982). The central problem in the epistemology of mathematical theories concerns an explanation of mathematical reality: What is it and how do we prove the propositional facts of mathematics (Kitcher, 1984; Nesher, 2002: X)? Since this reality cannot be known by any axiomatic mathematical theory, there may be other methods to know it, such as Gödel's mathematical intuition grasping mathematical true facts, or rather the epistemic logic we operate to quasi-prove the truth of our perceptual judgments representing mathematical reality (Agazzi, 1974: 24).

> (Assuming the consistency of classical mathematics) one can even give examples of propositions (and in fact of those type of Goldbach or Fermat) that, while contentually true, are unprovable in the formal system of classical mathematics. Therefore, if one adjoins the negation of such a proposition to the axioms of classical mathematics, one obtains a consistent system in which a contentually false proposition is provable. . . . (Gödel 1931a: 203).

The discrepancy between Gödel's intuition about the realistic nature of mathematics and his attempt to prove propositional facts formally can be resolved by the Peircean epistemic logic of complete proofs. Through it, we can prove the truth of the basic propositional facts of mathematics, discover hypothetical axioms, and evaluate their truth upon the true facts of mathematical reality.

The question is, why nevertheless did Gödel's formal proof of the incompleteness of mathematical theories were accepted almost without questioning the problematic "epistemological antinomy?" It may be that the generation of Frege and Hilbert, and the next one, were captivated by the deductivist-formalist agenda and the analytic formal semantic epistemology with the metalanguages hierarchies, which could not seriously reevaluate this proof (Dawson, 1984). Since the realistic conception of mathematics expresses mathematicians' intuition about their work, then what Gödel offered about the incompleteness of mathematical theories is accepted naturally: i.e., that there are "contentually true" propositions in the

language of theory that cannot be proved except by extended axiomatic theories (Hintikka, 2000: V).

5. The Pragmaticist Epistemology of Cognitive Empirical Representations of External Reality

The deviation of formal systems from human working with mathematical theories can be explained by suggesting that *formal systems are only realistic theories in disguise* or *utopian; i.e., impossibly "ideal machines*" of different degrees (Dawson, 1984:79; Nesher, 2001b).

> By the turn of this century mathematics, 'the paradigm of certainty and truth', seemed to be the real stronghold of orthodox Euclideans. But there are certainly some flaws in the Euclidean organization even of mathematics, and these flaws caused considerable unrest. Thus the central problem of all foundational schools was: 'to establish once and for all the certitude of mathematical methods'. [1] ([1] Hilbert, 1925). However, foundational studies unexpectedly led to the conclusion that a Euclidean reorganization of mathematics as a whole may be impossible; that at least the richest mathematical theories were, like scientific theories, quasi-empirical. Euclideanism suffered a defeat in its very stronghold (Lakatos, 1978: 30).

The formal systems with their formal proofs, though aiming to increase the power of formal computations, yet as far as they estranged from human cognitive operations representing reality their efficiency is decreased. The advantage of human cognitive operations lies in its having self-consciousness and self-control in confronting the mathematical, physical, and other realities, which enable correcting errors and evolving human knowledge (Gödel, 1972a: 305-6; Nesher, 1990, 1999; Hintikka, 1997: 5.7, 2000: X; Putnam, 2011: 15.4). In this perspective, we can understand the epistemology of the "Exact Sciences," the issue of the Königsberg Conference in September 1930, in which Gödel announced his discovery of incompleteness; namely, that even mathematics is not pure science and is only relatively exact (Nesher, 2002: X).

> . . . as far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. (Einstein, 1921)

Gödel's incompleteness theorem is about the relativity of any mathematical theory in respect to its proof-conditions in representing mathematical reality.

> There is in fact in the light of hindsight a major puzzle about Gödel's insights and about the way he put them to use. One of his greatest achievements, arguably the greatest one, was to show the deductive incompleteness of elementary arithmetic. (Hintikka, 2005: 536)

Hintikka obscures the issue that the incompleteness of any scientific theory, including elementary arithmetic, is due not only to the incompleteness of formal deductive inferences; scientific theories with their complete epistemic logical proofs are also incomplete and are true only upon their specific proof-conditions and therefore, they are incomplete in respect to reality we endeavor to represent. Since all our

knowledge of reality is based on perceptual experience in confrontation with reality, so also is our mathematical experience in confrontation with its reality, which cannot comprise Platonist abstract objects. The distinction between completeness of axiomatic formal systems and the incompleteness of mathematical and other scientific theories is not logical but, rather, epistemological and can be proved with pragmaticist epistemic logic (Nesher, 2002, 2007; Wittgenstein, 1933-34: 296).

> The nontriviality of the proof of completeness for limpid logic must be forcefully presented the possibility to Platonist Gödel that there were propositions that are *arithmetically* true but not provable within a formal system of arithmetic. (Goldstein, 2005: 154)

Thus, Gödel's "evident without proof" of true propositions that were not proved in specific formal systems, illustrates that cognitive confrontation with external reality cannot be formalized. According to Gödel the basic true mathematical facts can be grasped intuitively and from them the axioms are intuitively accepted as true without proofs.

> Of course, the task of axiomatizing mathematics proper differs from the usual conception of axiomatics insofar as the axioms are not arbitrary, but must be correct mathematical propositions, and moreover, evident without proof. There is no escaping the necessity of assuming some axioms or rules of inference as evident without proof, because the proofs must have some start point. (Gödel, 1951: 305).

However, since there is no human truths without proofs this can be undertaken only by quasi-proofs of basic perceptual judgments representing reality in complete epistemic logic, the trio sequence of the material logical inference of Abductive discovery, the Deductive necessary inference and the material inference of Inductive evaluation (Nesher, 2002: V, X). Hence, the impossibility of proving formally in metamathematics the theorem of unprovability is also due to the impossibility of proving formally the truth of propositional facts of external mathematical reality, "because the proofs must have some start point" and their proved truth is the "start point." This is hinted by Russell about the empirical assumptions of mathematics, and so Gödel, too, cannot prove G1 formally in an incomplete mathematical theory (Russell, 1914; Nesher, 2002: V). With the cognitive epistemic logic, we start from the quasi-proof of the basic perceptual facts of our knowledge of reality without any miraculous "given." Thus, we can discard the transcendental *a priorism* while all our knowledge is empirical (Nesher, 2007).

[3] The Entire Perceptual Operation: Complete *Trio* of Abduction, Deduction, and Induction:

Abduction$((C^{Ab}(A^{Ab}\rightarrow C^{Ab})\Rightarrow A^{Ab})$+**Deduction**$((A\rightarrow C)^{Ab}\ A^{Ab})\rightarrow C^{Dd})$+**Induction**$((A^{Ab}, C^{In})\sim\!>(A^{Ab}\rightarrow C^{In}))$

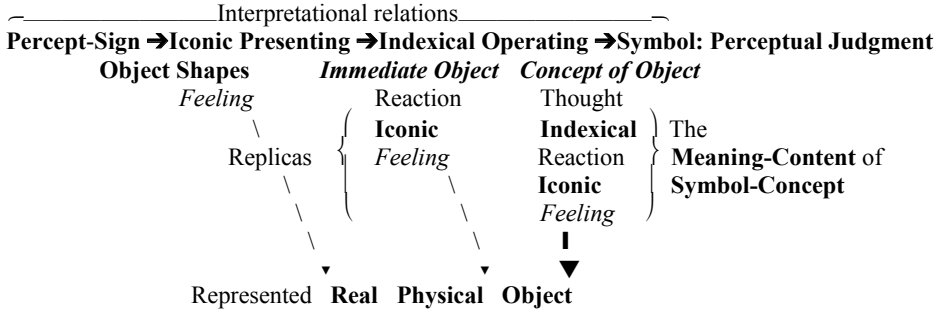
Where: => is the Abductive *plausibility connective* suggesting the concept A^{Ab}, → is the Deductive *necessity connective* from which the abstract object C^{Dd} is inferred, and $\sim\!>$ is the Inductive *probability*

connective evaluating the relationship between the concept A^{Ab} and the new experiential object C^{In}, when C^{In} is similar to C^{Dd}. From this epistemological position, it is amazing that Gödel, by using pure intuition and thus admitting the limitation of formal proofs, nevertheless attempted to prove the incompleteness of mathematical theories by incomplete formal inference (Gödel, 1931: #1, 1951: 304-306; Dawson, 1984: #2; Hintikka, 2005: 536).

Indeed, Lakatos and Putnam's conception of the *quasi-empirical* proofs in mathematics seem analogical to Gödel's mathematical proofs with intuitive grasps of true facts and his other intuitive inferences. Howevr, the Peircean epistemic logic of the *trio* inferences is the solution to the limitation of formal logic, yet not as the *quasi-empirical method* based on convention but *empirically quasi-proving* the *truth* of the basic propositions upon mathematical *external reality*. Thish is the only way to reach *convention* and for realism in human knowledge including mathematical knowledge (Lakatos, 1967[1978]: 36; Putnam, 1975: 63-77). The Pragmaticist overcoming of Gödel's Platonism is that all our knowledge develops from our sense-perception confrontation with external reality, and therefore conceptual realism with its pure intuition is only *disguised* empirical knowledge of reality. Since for Gödel mathematical reality consists of abstract entities, the analogy with empirical sciences is incomplete. The following is a schema of perceptual quasi-proof of perceptual judgment representing external reality (cf. [3]):

[4] **Perceptual Experience of Interpreting Cognitive Signs in Representing Physical Objects: Quasi-proof of the Truth of Perceptual Judgment**

I n t e r p r e t a t i o n relations evolve hierarchically
From Pre-verbal Sensorimotor Signs to Propositional Judgment



The signs representing a **Real Object** constitute the **Iconic** Feeling of **Object Shapes**, the **Indexical** reaction to it being the ***Immediate Object*** pre-symbolic representation, and their synthesis in the **Symbolic *Concept*** *represents* the ***Real Object*** by the true **Perceptual Judgment**. Recognizing that our knowledge

starts from perceptual confrontation with reality, we can understand Gödel's problem with *grasping* ideal entities through pure intuition, like the Kantian *Intellectual Intuition* in grasping *supersensible objects*, which only a *supernatural being* can do (Gödel, 1951; Dummett, 1981: 251-252). It is upon such basic knowledge that all our theories develop through the discovery of hypotheses (Nesher, 2008).

> But despite their remoteness from sense-experience, we do have something like a perception of the objects of set theory, as it seems from the fact that the axioms *force* themselves upon us as being true. I don't see any reason why we should have less confidence in this kind of perception, i.e., in mathematical intuition, than in sense-perception, which induces us to build up physical theories and to expect that future sense perceptions will agree with them, and, moreover, to believe that a question not decidable now has meaning and may be decided in the future. (Gödel, 1964: 268; emphasis added; Weyl, 1949: 235)

We can compare this feeling of *force* to Frege's feeling the *force* of truth in indicative sentences:

> We declare the recognition of truth in the form of an indicative sentence. We do not have to use the word "true" for this. And even when we do use it, *the real assertive force* lies not in it but in the form of the indicative sentence, and where this loses its assertive force the word "truth" cannot put it back again. (Frege, 1918: 89-90, emphasis added; cf. Nesher, 2002: VI.5.)

Such a feeling of the force of truth is the feeling of the self-controlled perceptual quasi-proofs of our perceptual judgments, and "the fact that the axioms *force* themselves upon us" is the feeling of the Abductive discovery and Inductive evaluation of the axioms as hypotheses, through the instinctive, practical and rational operation of epistemic logic. Thus, mathematical theories are also based on perceptual experience confronting its external reality. The question is how mathematical reality differs from physical reality (Putnam. 1975: #4, 1994: # 12).

6. What, Therefore, Is the *Mathematical Reality* That Mathematical Theories Represent?

Since all our knowledge of reality is based on perception and introspection, then basic mathematical knowledge is also based on such experiences (Wang, 1974: VII.3; Nesher, 2002: III). The basic *Mathematical reality* that we initially represent consists of *our operations of counting, grouping, and measuring physical objects* when confronting our environment (Nesher, 1990, 2002: V, 2007).

> . . . the primitive man could count only by pointing to the objects counted, one by one. Here the object is all-important, as was the case with early measures of all peoples. The habit is seen in the use of such units as the foot, ell (elbow), thumb (the basis for our inch), hand, span, barleycorn, and furlong (furrow long). In due time such terms lost their primitive meaning and we think of them as abstract measures. In the same way the primitive words used in counting were at first tied to concrete groups, but after thousands of years they entered the abstract stage in which the group almost ceases to be a factor. (Smith, 1923: 7)

Hence, arithmetic and geometry were historically basic human modes of quantitative operations on

physical objects. With our sensual perception, we represent these operations, yet not the engaged physical objects and not the involved conceptual number signs, but their combination in these operations themselves. Hence, the perceptual representation of these operations, being our basic representation of mathematical reality, is "a kind of visual justification which the Egyptian employed" (Gittleman, 1975: 8, 27-31; Parsons, 1995: 61). The arithmetical numbers are neither *physical objects* nor *abstract concepts*, but the *conceptual components of our quantitative operations with physical objects*. We assign numbers to these intentional cognitive operations *cum* physical maneuvers as signs of these operations. The *discovery* of the first concepts of these operations of enumeration consist of natural numbers; and the further *discovering* of their expansion through abstractions and generalizations constitutes our new mathematical hypotheses, which will be evaluated upon the extended mathematical reality (Gödel, 1944:128, 1964:268; Martin, 2005: 207; Spinoza, 1663).

> But consider a physical law, e.g., Newton's Law of Universal Gravitation. To say that this law is true . . . one has to quantify over such non-nominalistic entities as forces, masses, distances. Moreover, as I tried to show in my book, to account for what is usually called 'measurement' – that is, for the numericalization of forces, masses, and distances – one has to quantify not just over forces, masses, and distances construed as physical properties . . . , but also over *functions from* masses, distances etc. *to real* numbers, or at any rate to rational numbers. In short – and this is the insight that, in essence, Frege and Russell already had – a reasonable interpretation of the *application* of mathematics to the physical world *requires* a realistic interpretation of mathematics. (Putnam, 1975: 74)

The realistic understanding of mathematics that I suggest here is that mathematical reality is not an interpretation in the physical reality the physical sciences represent but it is the human operations of counting, groping, and measuring physical objects and their relations, being the basic mathematical reality upon its true representation the mathematical abstract and generalized theories are developed (Putnam, 1975: 77-78; Weyl, 1949: 235).

These basic operations are known by their perceptual representations; however, when we abstract, generalize, and further recombine the arithmetical components of these operations with our intellectual intuition, we continue to self-control them perceptually. Although the new mathematical structures are based on our perceptual confrontation with the reality of operations, when we elaborate them into more complicated kinds of mathematical structures they seemed detached from their reality as abstract conceptual entities grasped by pure intuition. Actually they are evolving in hierarchical relations between *sense-perception* and *intellectual intuitions* in our knowledge of mathematical reality without this reality being divided into "two separate worlds (the world of things and the world of concepts") (Gödel, 1951: 321).

> On the other hand, we have a debate between Realism—mathematical things exist objectively, independently of our mathematical activity—and Constructivism—mathematical things are created

> by our mathematical activity. We want to know how much of this can be regarded as continuous with the practice itself. (Maddy, 1997: 191)

The question is about the relationship of our mathematical activity with mathematical structures such that if they are external mathematical reality how we know them, and if they are our constructions, how can we apply them in our empirical theories (Heyting, 1931: 52-53; Dedekind, 1901:15-16)? The solution to this predicament between Metaphysical Realism and Phenomenological Constructivism is that mathematical reality *exists objectively,* yet *not independently* of our mathematical activity. Mathematical reality is our intentional self-controlled mathematical operations on physical objects, such as 1 apple and 1 apple are 2 apples, which are connected with our perceptual representation of this operation as a certain behavioral reality. Hence, we perceptually quasi-prove the truth of our perceptual judgment that "1 + 1 = 2," representing a mathematical operation, and thereby discover the structures of arithmetical numerical signs. Then, by discovering and proving the true representation of new mathematical operations, we hypothesize general theories, such as Peano's Arithmetic; finally, by evaluating them, we extend our knowledge of mathematical reality (Smith, P., 2007: #28.3). In this way we discover the construct of mathematical theories although the Constructivists consider the theories themselves as mathematical reality and not as representations of mathematical operations reality (Resnik, 1997). Hence, only by quasi-proving the truth of perceptual facts representing mathematical operations do we represent mathematical reality.

[5] The Double Layer of Mathematical Operations: (1) Counting Physical Objects; (2) Perceptual Quasi-proving the Truth of Discovering the Numerical Signs of the Operation (Peirce, 7.547)

I n t e r p r e t a t i o n Relations evolve From Pre-verbal Signs to Propositional Judgment

The Cognitive Representation of Mathematical Reality*: Discovering and Operating Numerical *Signs

⌐——————Reflective Interpretational Relations——————¬

(2) Percept-Sign➔Iconic Presenting➔Indexical Operating➔Symbolic Notion: Perceptual Judgment

Object Shapes ***Immediate Object*** Representing Reality **Numerical Counting**

(1) *Human Self-Controlling of Numerical Operations* of *Counting and Measuring Physical Objects*

⌙—————— **Mathematical Reality** ——————⌙

Gödel considers abstract mathematical theories analogous to physical theories such that mathematical axiomatic theories representation of mathematical abstract reality precedes their application to the empirical world but it is not the reality of human mathematical operations themselves on physical objects:

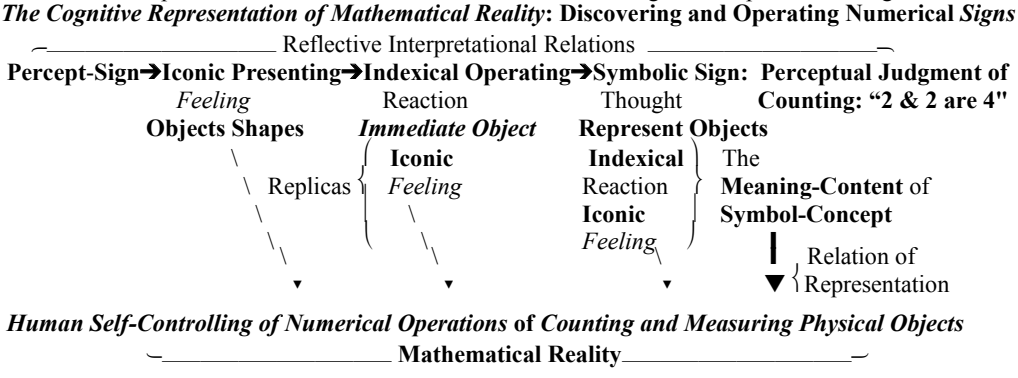
> ". . . the applications of mathematics to the empirical world, which formerly were based on the intuitive truth of the mathematical axioms, . . ." (Gödel, 1953:#12)

In contrast to Gödel's role of intuition to grasp the truth of mathematical abstract facts, we can perceptually prove the truth of propositional facts representing the reality of mathematical operations (Wittgenstein, 1956: III, 44). By understanding that mathematical reality consists of perceptually self-controlled operations, we can see how Gödel confuses the meaning-contents of mathematical symbols, which are the immediate modes representing numerical operations, with his Platonist mathematical abstract objects. These immediate modes of representation are the Peircean indexical representations of real objects which in mathematics are the factual operations of mathematical reality. Here we can discern Gödel's close insight of Peirce's conception of the perceptual "immediate object" component of symbols representing mathematical reality (Peirce, *CP*: 8.183, 8.343 [1908]; Nesher, 2002: II).

> It should be noted that mathematical intuition need not be conceived of as a faculty giving an *immediate* knowledge of the objects concerned. Rather it seems that, as in the case of physical experience, we *form* our ideas also of those objects on the basis of something else which is immediately given. Only this something else here is *not* or not primarily, the sensations. That something beside the sensations actually is immediately given follows (independently of mathematics) from the fact that even our ideas referring to physical objects contain constituents qualitatively different from sensations or mere combinations of sensations, e.g., the idea of object itself, whereas, on the other hand, by our thinking we cannot create any qualitatively new elements, but only | reproduce and combine those that are given. Evidently the "given" underlying mathematics is closely related to the abstract elements contained in our empirical ideas. It by no means follows, however, that the data of this second kind, because they cannot be associated with actions of certain things upon our sense organs, are something purely subjective, as Kant asserted. Rather they, too, may represent an aspect of objective reality, but, as opposed to the sensations, their presence in us may be due to another kind of relationship between ourselves and reality. (Gödel, 1964: 268)

Here Gödel's distinction between sensual perceptions and mathematical intuitions of the reality of abstract mathematical objects is the Pragmaticist distinction between the immediate iconic-sensual sign and the indexical-reaction being the "immediate object," the "abstract element" which is only the sign *representing* the *real object*. This Gödel's distinction is based on a confused epistemology that replaces the *meaning-contents* of such mathematical propositions with the *external reality they represent* (Gödel, 1953/54?: #35). It is Peirce's conception of the cognitive "immediate object," representing the real object that Descartes calls "objective reality" in distinction from "formal reality," the real object, without being able to explain it as perceptual cognitive representation of external reality (e.g., Peirce, CP: 8.183, 8.343; Nesher, 2002: II, III, V; Feferman, 1998; Parsons, 2008: Chap. 6). The following is a schema of a mathematical reality operation represented by the perceptual *immediate object* as the meaning-content of the symbolic sign of mathematics:

[6] Perceptual Representation of the Cognitive Operation of Counting Physical Objects by Quasi-proving the Truth of Its Perceptual Judgment of Mathematical Operation



An echo of this explanation is noticed in Gödel's insight into the realist nature of mathematics:

> . . . [mathematics] in its simplest form, when the axiomatic method is applied, not to some hypothetico-deductive system as geometry (where the mathematician can assert only the conditional truth of the theorems), but mathematical proper, that is, to the body of those mathematical propositions, which hold in an absolute sense, without any further hypothesis. There must exist propositions of this kind, because otherwise there could not exist any hypothetical theorems | either. For example, *some* implications of the form:
>
> > If such and such axioms are assumed, then such and such theorems hold, must necessarily be true in the absolute sense. Similarly, any theorem of finitistic number theory, such as 2 + 2 = 4, is no doubt, of this kind. (Gödel, 1951: 305; cf. 322)

The perceptual representation of basic mathematical operations is the quasi-proved true empirical facts of mathematical reality, but not an ideal one. Yet this seems to be an unbridgeable gap for Penrose.

> . . . real numbers are called 'real' because they seem to provide the magnitudes needed for the measurement of distance, angle, time, energy, temperature, or of numerous other geometrical and physical quantities. However, the relationship between the abstractly defined 'real' numbers and the physical quantities is not as clear-cut as one might imagine. Real numbers refer to *mathematical idealization* rather than to any actual physically objective quantity. (Penrose, 1989: 112-113; cf. Penrose, 2011: 16:1)

Hence, Popper's amazement as to why mathematics can be applicable to reality is resolved by explaining that mathematics indeed originated in human perceptual true representations of mathematical reality, the "empirical basis" of mathematical theory being more abstract component of this empirical science (Popper, 1963: #9; Dedekind, 1901: 17; Poincare, 1902: Author's Preface, Chap. II).

7. Mathematics Is an Empirical Science Based on True Propositional Facts of Mathematical Reality

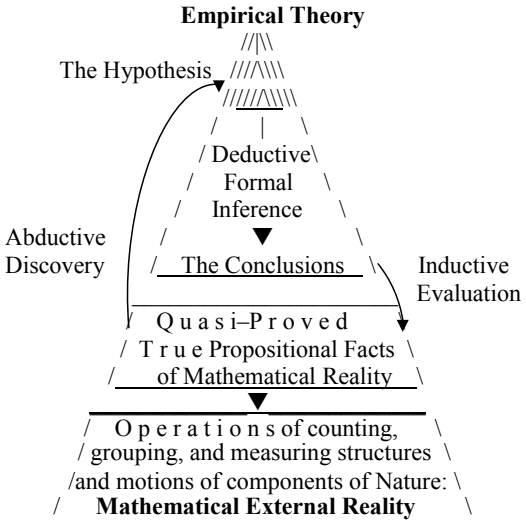
Hence the problem is to explain the nature of mathematical science and what are the "data," the basic facts upon them the mathematical theories develop and evaluated?

> . . . mathematics has always presented itself, throughout the history, as an abstract discipline, but has nevertheless always dealt with specific subject matter of its own. Considering mathematics in this light one might ask: what kind of knowledge can be attained through it? How can it be said to deal with contents and objects which are offered as 'data,' and yet are not data at all from the point of view of sensible experience? We are here confronted with the problem of mathematical intuition, considered as a real source of knowledge, to be clearly distinguished from that further form of mathematical activity which consists in the systematic construction of various theories. Indeed, the most delicate point of this problem is precisely the comparison between the intuitive moment and the moment of theoretical construction, since it is impossible to deny that, in many cases at least, mathematical theories are in fact an exact and systematic codification of what is known intuitively, and that, on the other hand, intuition is not sufficiently reliable unless it is supported by logical proof (Agazzi, 1974: 9-10).

The formal logical proof cannot support or replace the intuitive grasp of the mathematical basic true fact in Gödelian Platonism, and only the epistemic logic of Peircean trio can quasi-prove the truth of the perceptual judgments as the basic mathematical propositional facts (Nesher, 2002: X). Only this logic can replace the mysterious unexplainable intuition of mathematical facts and can prove mathematical truths by the epistemic complete proof. Thus it also replaces the assuming roles of such intuition for discovery and evaluation of the axioms of mathematical theories (Agazzi, 1974: 12).

From the quasi-proof of the truth of the basic mathematical propositional facts of mathematical reality, the mathematical hypotheses are Abductively *discovered* to infer Deductively their *predicted* theorems and *evaluated* Inductively upon empirically newly discovered and proved mathematical facts. The following is a pragmaticist epistemological explanation of the general structure and operation of the theories of mathematical empirical science:

[7] Pragmaticist Epistemological Presentation of Mathematical Empirical Theory:



The *proof-conditions* of **mathematical empirical theory** are the *epistemic logic*, the *trio* comprising *inferential rules of the complete proof* of the truth of basic *propositional facts* representing **external reality**. With this epistemic logic we also prove the truth of scientific hypotheses (Gödel's axioms), through their Abductive discovery, Deductive formally inferred theorems and their Inductive evaluation upon the basic *propositional facts*. Yet, Gödel's conception of mathematical intuition covers those different components of the Pragmaticist epistemic logic which though he felt their operations but could not explain the truth of these basic propositional facts of mathematical reality and the truth of the axioms which the epistemic logical complete proof can do (Feferman, 1998: #1; Parsons, 2008: #5). Hence, empirical theories are only relatively true by being "closed" upon their proof-conditions, which can change with newly discovered facts of reality (Heisenberg, 1971:43-44; Nesher, 2002: V.5, X.10).

> Yet if mathematical facts are facts, they must be facts about something; if mathematical truths are true, something must make them true. Thus arises the first important question: what is mathematics about? If 2 plus 2 is so definitely 4, what is it that makes it so? (Maddy, 1990:1)

Although mathematical theory is about mathematical operations of counting, grouping, measuring, and so on, the question is, how do we prove the mathematical facts representing such operations; i.e., "what is it that makes it so" that 2 plus 2 are definitely 4? We operate in such a manner that we count with our indexical ostensions while representing this operation in our perceptual judgment as a true fact of such arithmetical counting. Since all our basic knowledge comprises such quasi-proofs of our perceptual judgments, so too do the truths of our basic mathematical facts represent such operations of mathematical reality (comp. Hempel, 1945).

Indeed, we do not create on our will the patterns of mathematical reality, but we discover the mathematical concepts of our counting, grouping, and measuring operations with physical objects in the operations of mathematical reality, and this is "[mathematics] in its simplest form, . . . mathematical proper, that is, to the body of those mathematical propositions, which hold in an absolute sense, without any further hypothesis" (Gödel, 1951: 305; Dedekind, 1901:15-16). Epistemologically we can understand that when we intuit the force of the truth of our basic mathematical propositions we feel that they "hold in an absolute sense" but without conceiving the epistemic logic we cannot explain them as our own empirically quasi-proved true mathematical propositions (Steiner, 2000: 337-339).

> Namely, it is correct that a mathematical proposition says nothing about the physical or psychical reality existing in space and time, because it is true already owing to the meaning of the terms occurring in it, irrespectively of the world of real things (Gödel, 1951: 320).

Yet Gödel is right that mathematical reality consists of neither physical nor psychical realities but it is the specific connection between them; namely, the mathematical "world of real things" is our cognitive operations of quantifying components of physical reality, and the meaning-contents of mathematical signs evolve in this perceptual experience (Wittgenstein, 1956: III, 44; Benacerraf, 1973; Tait, 1986; Resnik, 1992: #1; Martine, 2005: 210).

> To mention another example, the Pitta-Pitta, a tribe [of aborigines] in Queensland, are able to count the fingers and toes without a system of numerals, but only by the aid of marks in the sand. . . (Smith, D., 1923: 7; Gullberg, 1997: Ch. 4).

This is evidence of arithmetical facts that are iconic cum indexical sensori-motoric operations of counting and grouping with pre-conceptual signs of properties and relations that eventually develop into conceptual components, the numerical symbols involving in mathematical facts (Gödel, 1951: 320).

> From its earliest beginnings science has used mathematics. Counting, measuring, ordering, and estimating are basic mental operations necessary for science as well as for many other human activities, and their nature is mathematical (Bos, 1993: 165).

Hence, mathematics, from "the ubiquitous use of elementary mathematics" to "the great variety of high level applications of mathematics" (Bos, 1993: 165-166), is an empirical science of the operational quantification of physical components of nature. Its development is from the use of elementary to the variety of high level mathematics evolved from the elaboration of abstract mathematical theories related to their advance applications by scientists working toward the advancement of scientific theories.

8. Conclusion: Mathematics Is an Empirical Science Representing Its Own Reality, Being Neither Queen Nor Servant of Other Empirical Sciences but Their Quantitative Backbone

The problem is to explain the difference between mathematical science and other sciences and their collaboration, when all are empirical sciences representing different realities and with different roles in developing our knowledge of nature (Wang, 1974: VII). Thus, in mathematics we cannot have true theories without proving them upon mathematical reality. Mathematicians develop their theories by discovering general hypotheses as mathematical formulations of theoretical models, typically of physics, like of fields of forces and topology of fluid flows, but of all other sciences, and evaluate them upon mathematical reality of quantitative operations on predicted physical observations.

> The rich interplay between mathematics and physics predates even their recognition as separate subjects. The mathematical work that in some sense straddles the boundaries between the two is commonly referred to as *mathematical physics*, though a precise definition is probably impossible. (Jaffe & Quinn, 1993: 4)

Mathematical theories formularize models for theoretical physical hypotheses, but there is a

distinction between proving the truth of mathematical theories and proving the truth of the relevant physical theories themselves (Feferman, 1998).

> Fore as far as verifiable consequences of theories are concerned the mathematical axioms are exactly as necessary for obtaining them as the laws of nature (cnf. footn. 41). If, e.g., the impredicative axioms of analysis are necessary for the solution of some problem of mathematical physics, these axioms will imply predictions about observable facts not obtainable without them. Moreover it is perfectly conceivable that an inconsistency with observation may be due to not to some wrong physical assumptions but to an inconsistency of these axioms. (Gödel, 1953-54 II: #44, p. 188)
> That it is arbitrary to call mathematics void of content because, without laws of nature, it has no verifiable consequences also appears from the fact that the same is true for the laws of nature without mathematics or logic. cf. also #44. (Gödel, 1953-54 II: fn. 41, p. 207)

Thus, physicists and mathematicians have different realities to represent with their theories, and the mathematical theory which proved true in the measurement of observed physical facts is only the condition for the evaluation of physical theories. Thus, in distinction from Gödel's conceptual epistemology of mathematics, according to the above explanation, the mathematical reality is also empirical. The truth of mathematical theory enables proving experimentally the truth but also the falsity of physical theories. In this way, we can understand the Gödelian epistemic intuition about the nature of mathematical theories, yet not the Quinean "mathematical naturalism," which confuses mathematics with other sciences and identifies mathematical reality with physical reality.

When there are difficulties with a physical picture of reality and the mathematical model for it, such that it becomes impossible to make measurable predictions, then the problem is to inquire what is wrong that we are unable to evaluate experimentally the physical hypothesis (Woit, 2007: x-xiii, Ch. 14; Feferman, 1998: #2, #4).

> I can't say whether string theory will ever get past its most serious hurdle–coming up with a testable prediction and then showing that the theory actually gives us the right answer. (The math part of things, as I have said, is already on a much firmer ground.) Nevertheless, I do believe the best chance for arriving at a successful theory lies in pooling the resources of mathematicians and physicists, combining the strengths of the two disciplines and their different ways of approaching the world. (Yau & Nadis, 2010: 304)

Hence, mathematics without operational measuring the predicted and eventually observed true facts of reality cannot be true and cannot be "on a much firmer ground" than physics without "a testable prediction." Both have to prove their own truths upon "their different ways of approaching the world."

> However mathematical intuition in addition creates the conviction that, if these formulas express observable facts and were obtained by applying mathematics to verified physical laws (or if they express ascertainable mathematical facts), then these facts will be brought out by observation (or computation) (Gödel, 1953/9-III: #16; cf. ##13-15 & n. 34).

How may one understand this hinted explication of the relationship between intuitive mathematical truth representing its own reality and its application to physical theories to enable observable predictions of them (Gödel, 1953II: #15)? In the end, mathematics is neither the *queen* of science nor its *servant* but its *quantitative backbone*—that is, the quantified formulations of scientific theoretical models and their operations on scientific observations—without which physical and other theories cannot be evaluated experimentally (Bos, 1993: #10). The explanation to the puzzlement why mathematics is considered *exact* or *pure* science while being empirical like other experimental sciences, is the relative simplicity of its represented reality in respect to the physical and the psychological realities.

> Mathematics may be the queen of the science and therefore entitled to royal prerogatives, but the queen who loses touch with her subjects may lose support and even be deprived of her realm. Mathematicians may like to rise into the clouds of abstract thought, but they should, and indeed they must, return to earth for nourishing food or else die of mental starvation. They are on safer and saner ground when they stay close to nature. (Kline, 1959: 475)

This is a poetic metaphor that illustrates the above explanation of the empirical nature of mathematical reality, upon which mathematical theories can be evaluated and be proved true. This empirical explanation can be seen in Gödel's late philosophical writings on the foundations of mathematics:

> If mathematics describes an objective world just like physics, there is no reason why inductive methods should not be applied in mathematics just the same as in physics. . . . This whole consideration incidentally shows that the philosophical implications of the mathematical facts explained do not lie entirely on the side of rationalistic or idealistic philosophy, but that in one respect they favor the empiricist viewpoint. It is true that only the second alternative points in this direction. (Gödel, 1951: 313)

Hence, we can know experientially the mathematical facts of the mathematical empirical reality.

References:

Agazzi, A. (1974) "The Rise of Foundational Research in Mathematics." *Synthese*, Vol. 27 Nos. 1/2 1974: 7-26.

Benacerraf, P. (1973) "Mathematical Truth." In W.D. Hart, ed., (1996) *The Philosophy of Mathematic.*

Benacerraf, P. and H. Putnam, eds. (1964) *Philosophy of Mathematics: Selected Readings*, Second Ed. Cambridge, UK: Cambridge University Press.

Berto, F. (2009) *There's Something About Gödel: The Complete Guide to the Incompleteness Theorem*. Oxford: Wiley Blackwell.

Bos, H.J.M. (1993) *Lectures in the History of Mathematics*. Providence, RI: American Mathematical Society.

Chihara, C.S. (1982) "A Gödelian Thesis Regarding Mathematical Objects: Do They Exist? and Can We Perceive Them?" *The Philosophical Review*, XCI, No. 2 (April) 1982.

Dawson, J.W. (1984) "The Reception of Gödel's Incompleteness Theorems." *The Philosophy of Science Association*, 1984, Vol. 2: 253-271. Reprinted in S.G. Shanker, ed., 1988: 74-95.

Dedekind, R. (1901) *Essays on the Theory of Numbers*. General Books 2009.

Devlin, K. (2002) "Kurt Gödel—Separating Truth from Proof in Mathematics." *Science* 6, December 2002, Vol. 298, No. 5600:1899-1900.

Dummett, M. (1981) *Frege Philosophy of Language*. Cambridge, MA: Harvard University Press.
Einstein, A. (1921) "Geometry and Experience," delivered to the Prussian Academy of Science, 1921.
Feferman, S. (1984) "Kurt Gödel: Conviction and Caution." In S.G. Shanker, ed., 1988: 96-114.
_____ (1998) "Mathematical Intuition vs. Mathematical Monsters." Expended version of a paper presented in the 20th World Congress of Philosophy, Boston MA, August, 8-11, 1998.
_____ (2006) "Are there Absolutely Unprovable Problems? Gödel's Dichotomy. " *Philosophia Mathematica* (III) 2006:1-19,
Frege, G. (1918) "The Thought: A Logical Inquiry." In Blackburn and Simmons, eds., *Truth*. Oxford: Oxford University Press: 85-195.
Floyd, J. (2001) "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth." *Philosophia Mathematica* (3) Vol. 9, 2000:280-307.
Floyd J. and H. Putnam (2000) "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel Theorem." *The Journal of Philosophy*, 2000:624-632.
Franzèn, T. (2005) *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. Natick, MA: A.K. Peters.
Frege, G. (1918) "The Thought: A Logical Inquiry." In S. Blackburn & K. Simmons, eds., *Truth*, 85-105. Oxford: Oxford University Press.
Gittleman, A. (1975) *History of Mathematics*. Columbus, OH: Merrill Pub. Comp.
Goldstein, R. (2005) *The Proof and Paradox of Kurt* Gödel. New York: W.W. Norton & Company.
Gödel, K. (1929) "On the Completeness of the Calculus of Logic." *Kurt Gödel Collected Works*, Vol. I, 1986: 61-101.
_____ (1931) "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems." *Kurt Gödel Collected Works*, Vol. I, 1986: 196-199.
_____ (1931a) "Discussion on Providing a Foundation for Mathematics." *Kurt Gödel Collected Works*, Vol. I, 1986: 201-204.
_____ (1934) "On Undecidable Propositions of Formal Mathematical Systems." *Kurt Gödel Collected Works*, Vol. I, 1986: 346-371.
_____ (1944) "Russell's Mathematical Logic." *Kurt Gödel Collected Works*, Vol. II, 1990: 119-141.
_____ (1946) "Remarks before the Princeton Bicentennial Conference on Problems in Mathematics." *Kurt Gödel Collected Works*, Vol. III, 1995: 150-153.
_____ (1947) "What is Cantor's Continuum Problem." *Kurt Gödel Collected Works*, Vol. II, 1990: 176-187.
_____ (1951) "Some Basic Theorems on the Foundations of Mathematics and Their Implications." *Kurt Gödel Collected Works*, Vol. III, 1995: 304-323.
_____ (1953-54?) "Is Mathematics Syntax of Language?, II." In: F.A. Rodriguez-Consuegra, ed., *Kurt* Gödel*: Unpublished Philosophical Essays*. Basel: Brikhäuser-Verlag, 1995, 1: 710-211.
_____ (1964) "What is Cantor's Continuum Problem." *Kurt Gödel Collected Works*, Vol. II, 1990: 254-270.
_____ (1972a) "Some Remarks on the Undecidable Results." *Kurt Gödel Collected Works*, Vol. II, 1990: 305-306.
Gullberg, J. (1997) *Mathematics from the Birth of Numbers*. New York: W.W. Norton.
Hart, W.D., ed. (1996) *The Philosophy of Mathematics*. Oxford: Oxford University Press.
Heisenberg, W. (1971) *Across the Frontiers*. New York: Harper Torchbooks, 1974.
Hempel, C. G. (1945) "On the Nature of Mathematical Truth." *The American Mathematical Monthly*, vol. 52 (1945): 543-556.
Heyting, A. (1931) "The Intuitionist Foundations of Mathematics." In: P. Benacerraf and H. Putnam, eds., (1964): 52-61.
Hilbert, D. (1925) "On the Infinite." Translated into English in J. van Heijenhoort (ed.) *From Frege to Gödel*. Pp. 367-92. Harvard University Press.
Hintikka, J. (1997) *Lingua Universalis vs. Calculus Ratiocinator: An Ultimate Presupposition of Twentieth-Century Philosophy*. Selected Papers 2. Boston: Kluwer Academic Pub.
_____ (2000) *On Gödel*. Wadsworth Philosophical Series.
_____ (2005) "What Platonism? Reflections on the Thought of Kurt Gödel." *Revue internationale*

de philosophie: 2005/4.
Jaffe, A. and F. Quinn (1993) "'Theoretical Mathematics': Toward a Cultural Synthesis of Mathematics and Theoretical Physics." *Bulletin of the American Mathematical Society*. Vol. 29, No.1 1993:1-13
Kitcher, P, (1984) *The Nature of Mathematical Knowledge*. Oxford: Oxford University Press.
Kleene, S.C. (1967) *Mathematical Logic*, Mineola, New York: Dover Publications.
Kline, M. (1959) *Mathematics and the Physical World*. New York: Dover Pub., 1981.
Lakatos, I. (1967) "A Renaissance of Empiricism in the Recent Philosophy of Mathematics." Chap. 2 in I. Lakatos, *Mathematics, Science and Epistemology: Philosophical papers Volume 2*. Edited by J. Worrall and G. Currie. Cambridge: Cambridge University Press 1978.
Maddy, P. (1990) *Realism in Mathematics*. Oxford: Clarendon Press.
_____ (1997) *Naturalism in Mathematics*. Oxford: Clarendon Press: Chapter 4.
Martin, D.A. (2005) "Gödel's Conceptual Realism." *The Bulletin of Symbolic Logic*, Vol. 11, No. 2 (June 2005): 207-224.
Nesher, D. (1987) "Epistemological Investigations: Is Metalanguage Possible? Evolutionary Hierarchy vs. Logical Hierarchy of Language." In *Development in Epistemology and Philosophy of Science*, Holder-Pichler-Tempsky 1987:72-80.
_____ (1990) "Understanding Sign Semiosis as Cognition and as Self-conscious Process: A Reconstruction of Some Basic Conceptions in Peirce's Semiotics." *Semiotica*, Vol. 79, Nos. 1/2: (1990):1-49.
_____ (1992) "Wittgenstein on Language and Reality: Meaning and Truth." The 15th Wittgenstein International Symposium, Kirchberg, August 1992.
_____ (1999) "Peirce's Theory of Signs and the Nature o0f Learning Theory." M. Shapiro, ed., *The Peirce Seminar Papers: Essays in Semiotic Analysis*, Vol. IV. New York: Berghahn Books: 349-388.
_____ (2001a) "Peircean Epistemology of Learning and the Function of Abduction as the Logic of Discovery." *Transactions of the Charles S. Peirce Society*, 2001, Vol. XXXVIII, No. 1/2: 175-206.
_____ (2001b) "The Three Languages of the *Tractatus*, Ordinary, Philosophical and of Natural Science: Can Wittgenstein Explain Our Representation of Reality?" for the Conference Wittgenstein Today. Bologna December 20-22, 2001.
_____ (2002) *On Truth and the Representation of Reality*. New York: University Press of America.
_____ (2007) "How to Square (*normo*, CP:2.7) Peirceanly the Kantian Circularity in the Epistemology of Aesthetic as Normative Science of Creating and Evaluating the Beauty of Artworks." Presented in the conference on Charles Sanders Peirce's Normative Thought. June, 2007, University of Opole Poland.
_____ (2008) "What Is Genius? A Pragmaticist Criticism of the Kantian Dichotomy between Artist–Genius' Productive Imagination Creating Freely Fine Arts and Scientist Following Mechanical Rules Determine the Formation of Theories." Delivered in the Sixth Quadrennial Fellows Conference of the Center for Philosophy of Science, Sunday, 20 July to Thursday, 24 July 2008, at Ohio University, Athens, Ohio.
_____ (2010) "The Role of Productive Imagination in Creating Artworks and Discovering Scientific Hypotheses." Papers of the 33rd International Wittgenstein Symposium: Image and Imagining in Philosophy, Science, and the Arts. Vol. XVIII. Kirchberg, Austria, 8-14 August 2009.
Parsons, C. (1995) "Platonism and Mathematical Intuition in Kurt Gödel's Thought." *Bulletin of Symbolic Logic*, Vol. 1, No. 1 (March 1995): 44-74.
Peirce, C.S. (1931-1958) *Collected Papers*, Vols. I-VIII, Harvard University Press. [*CP*]
_____ (2008) *Mathematical Thought and Its Objects*. Cambridge: Cambridge University Press.
Penrose, R. (1989) *The Emperor's New Mind*. Oxford: Oxford University Press.
_____ (2011) "Gödel, the Mind, and the Laws of Physics." In M. Baaz et al, eds., Kurt *Gödel and the Foundations of Mathematics: Horizons of Truth*. Cambridge: Cambridge University Press: 339-358.
Poincaré, H. (1902) *The Value of Science*, Chap. I: "On the Nature of Mathematical Reasoning." New

York: The Modern Library.
Popper, K. (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. Oxford: Clarendon Press.
Putnam, H. (1975) "What is Mathematical Truth?" In H. Putnam, *Mathematics Matter and Method*. Cambridge: Cambridge University press 1975: chap. 4.
_____ (1994) *Words and Life*. Ed. J. Conant, Cambridge Mass.: Harvard University Press.
_____ (2011) "The Gödel Theorem and Human Nature." In M. Baaz et al, eds., Kurt *Gödel and the Foundations of Mathematics: Horizons of Truth*. Cambridge: Cambridge University Press: 325-357.
Resnik, M.D. (1992) "Proof as the Source of Truth." In M. Detlefsen, ed., *Proof and Knowledge in Mathematics*. London: Routledge, 1992*:* 6-32.
_____ (1997) *Mathematics as a Science of Patterns*. Oxford: Oxford University Press.
Rodriguez-Consuegra, F. A. (1995) "Introduction," In F. A. Rodriguez-Consuegra, ed., *Kurt Gödel Unpublished Philosophical Essays*. . Basel: Brikhäuser Verlag: 24.
Russell, B. (1914) *Our Knowledge of the External World*. London: Routledge, 1993.
Smith, D.E. (1923) *History of Mathematics*, Vol. I. New York: Dover Pubs. 1951.
Smith, P. (2007) *An Introduction to* Gödel's *Theorems*. Cambridge: Cambridge University Press.
Shanker, S.G., ed., (1988) *Gödel's Theorem in Focus*. London: Routledge.
Spinoza, B. (1663) Letter 12 to Ludewijk Meyer. In *The Collected Works of Spinoza*, Ed. and Trans. by Edwin Curley. Princeton: Princeton University Press 1985.
Steiner, M. (2000) "Mathematical Intuition and Physical Intuition in Wittgenstein's Later Philosophy." *Synthese* 125: 333-340, 2000.
Tait, W.W. (1986) "Truth and Proof: The Platonism of Mathematics." In Hart, W.D., ed., *The Philosophy of Mathematics*. Oxford: Oxford University Press, 1996: 142-167.
Tarski, A. (1933) "The Concept of Truth in Formalized Languages." In A. Tarski, *Logic, Semantics, Metamathematics*. J.H. Woodger, trans. Oxford: Clarendon Press, 1956: 152-277.
_____ (1944) "The Semantic Conception of Truth." In H. Feigl & W. Sellars, eds., *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts, 1949: 52-84.
Wang, H. (1974) *From Mathematics to Philosophy*. London: Routledge and Kegan Paul.
Weyl, H. (1925) "The Current Epistemological Situation in Mathematics." In P. Mancosu, ed., *From Brouwer to Hilbert*. Oxford: Oxford University Press, 1998: 123-42.
_____ (1949) *Philosophy of Mathematics and Natural Science*. New York: Atheneum, 1963.
Zach, R. (2003) "Hilbert's Program." In *Stanford Encyclopedia of Philosophy*, 2003.
Wittgenstein, L. (1921) *Tractatus Logico-Philosophicous*. London: Routledge & Kegan Paul, 1961.
_____ (1933-34) *Philosophical Grammar*. Ed. R. Rhees, tr. A. Kenny. Oxford: Basil Blackwell, 1969.
_____ (1956) *Remarks on the Foundations of Mathematics*, G.H. von Wright et al., eds. Cambridge, MA: MIT Press, 1967.

Einstein as the Greatest of the Nineteenth Century Physicists

John D. Norton
Department of History and Philosophy of Science
Center for Philosophy of Science
University of Pittsburgh
http://www.pitt.edu/~jdnorton

This text is based on the chapter of the same name in my online textbook, *Einstein for Everyone* at http://www.pitt.edu/~jdnorton/teaching/HPS_0410/chapters/index.html

Modern day writers often endow Einstein with a 21st century prescience about physical theory that, it just so happens, is only now vindicated by the latest results of the same writers' research. There is a second side to Einstein. His science, methods and outlook were also clearly rooted in 19th century physics.

1. The Young and the Old Einstein

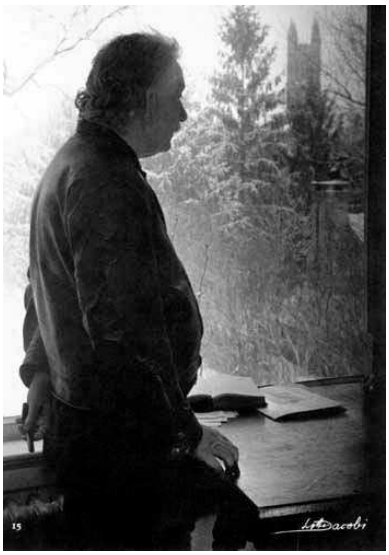
The Einstein of popular thought is the young Einstein. This is the intellectual rebel of 1905 who, in one year, laid out the special theory of relativity and $E=mc^2$, postulated the light quantum and used Brownian motion to make the case for the reality of atoms. These achievements were made prior to Einstein holding an academic position. He was then still a patent examiner in the Bern patent office. The years that followed brought Einstein a succession of ever more prestigious academic appointments; and, in the mid 1910s, he delivered his masterpiece, the general theory of relativity.

In all this, there was a real sense that Einstein was ahead of his peers, leading the way. The special theory of relativity was absorbed into the mainstream of physics fairly quickly. The general theory of relativity was

not quite so readily accommodated. This was in part due to its burdensome mathematical demands of the theory, at least relative to the standards of mathematical expertise then found among physicists. But the tide was flowing with Einstein. When the eclipse expeditions of 1919 vindicated Einstein's theory and he became a popular hero, critics risked being seen as unimaginative reactionaries.

Einstein's work on the light quantum did not fare so well. It was regarded by many as an odd aberration from an otherwise brilliant mind. Even in the early 1920s, it was doubted by Niels Bohr, who had a decade before developed the first quantum model of the atom.

By the end of the 1920s, however, another Einstein began to emerge. As the quantum theory enjoyed success after success, Einstein found himself unconvinced. He took on the role of critic, complaining that the new quantum theory, for all its virtues, could not be the final theory. This was Einstein's new place in the physics community for his final quarter century, ending with his death in 1955. He remained a revered figure. But he became increasingly isolated and marginalized, as he labored on his alternative theories with the help of a few assistants. In the years after his death, it became clear that Einstein's objection to quantum theory failed, but not, I believe, for the reasons articulated by his arch antagonist Niels Bohr.



The old Einstein is a recalcitrant Einstein, unwilling to swim with the new quantum tide that flooded over physics. We should not judge that harshly. No thinker can ever think purely new thoughts. We all sit at the junction of the old and the new. Einstein was one of the first of new physicists of the twentieth century. His discoveries and methods exercised a profound, defining influence on the development of twentieth century physics. However, there is also a strong sense in which he was one of the last of the nineteenth century physicists. Perhaps he was the greatest of them.

2. Themes of Nineteenth Century Physics

To see why this is not such an unreasonable assessment, we should review the major discoveries and themes of the nineteenth century and then see how they came to be realized and even fulfilled in Einstein's research. We shall look at three categories: Einstein's science, his methods and his outlook.

2.1 Science

Nineteenth Century...

Electrodynamics

The great discovery in physics of the nineteenth century was Maxwell's electrodynamics and its development by later physicists, including H. A. Lorentz.

Newton's physics had been very successful in recovering the properties of things like apples and planets that move at ordinary speeds, much less than that of light.

Nineteenth century electrodynamics now succeeded in probing the properties of things moving close to the speed of light: charges accelerated to high speeds and light itself.

Mixed in were now familiar kinematical effects: rapidly moving systems shrink in length and slow in time; and there are temporal dislocations over space. These were encoded in mathematical transformation equations discovered by H. A. Lorentz and others.

Einstein...

Special relativity

Einstein took special note of the kinematical effects mixed in with the electrodynamical effects of Lorentz's theory. He saw that they could be separated out as a novel theory of space and time, independent of the electrodynamics. The result was the special theory of relativity. Its central equations were the same transformation equations that Lorentz had employed in his development of electrodynamics.

Einstein is commonly understood as repudiating nineteenth century electrodynamics in his rejection of its ether. That assessment is altogether too narrow. In extracting the kinematics as an independent theory, Einstein was harvesting one of the greatest fruits of the nineteenth century theory. The ether was merely surplus foliage that needed to be trimmed away during the harvest.

The special theory of relativity is the natural completion of nineteenth century electrodynamic theory.

Nineteenth Century...

Thermal and statistical physics

Another significant achievement of nineteenth century physics was the final recognition that thermal processes were to be understood statistically, as the average behavior of systems made of very many components.

The simplest case was ordinary matter. It is made of atoms and molecules and heat resides in the energy distributed randomly over them.

The same analysis could be given of heat radiation. The many components are the many frequencies that comprise radiation.

In all cases, equilibrium thermal systems arise when energy is distributed in its most probable configuration over these components. Probabilities arise merely out of our ignorance of the precise microstate of the system.

Einstein...

Reality of atoms

When Einstein began work on thermal physics, this statistical approach was still struggling for mainstream acceptance. Boltzmann had shown that a molecular theory would fit with the known thermal properties of matter.

However he had failed to convince a significant portion of his community that they ought to adopt his molecular approach. Thermodynamicists, such as Nernst, had found their thermodynamic theories adequate to all observed thermal processes. Why should they trouble themselves with molecules too small to be seen?

If that attitude puzzles you, note that it is merely the analog of the electrodynamicist who is quite happy with the theory of electric and magnetic fields and resists ether theories that account further for these fields in terms of hidden ether machinery--little wheels spinning and meshing at microscopic scales in the ether.

Einstein's work of 1905 on Brownian motion was a major advance, perhaps even the major advance, that made acceptance of atoms inevitable. He showed that there were thermal phenomena that could ONLY be understood by Boltzmann's statistical methods. The thermodynamicists now had to adopt these methods; and they did.

2.2 Methods

Nineteenth Century...

Revealing the hidden mechanism

Einstein's earliest papers, starting in 1901, were devoted to discovering the discrete molecular mechanisms that underlie the continuity of thermal appearances. That goal is a nineteenth century one: completing the molecular program of Boltzmann. This was work that Einstein largely abandoned. However, during this early work, he developed methods for inferring from the observable thermal properties of substances to their microscopic structure.

The easiest and simplest of these was the observation of the ideal gas law in the thermal appearances. Robust argumentation showed that it must derive from a microstructure of very many, spatially localized components.

Systems of this simplest type were the ones Einstein investigated in his *annus mirabilis* of 1905. The tiny particles of Brownian motion form such a system. So do the sugar molecules in dilute solution investigated in his 1905 doctoral dissertation.

Einstein...

Light quantum

When it comes to Einstein's boldest posit, the light quantum, it is easy to find a prescient Einstein, somehow anticipating all the quantum craziness to come. In 1905, in a letter to a friend, Einstein was already calling this one result, among all those of 1905, "very revolutionary."

The result was properly called revolutionary when set in the context of electrodynamical theory. For it contradicted the picture of light as a wave whose energy is spread out over space.

There is another way to see it, as I reported in the chapter "Atoms and the Quanta." Set in the context of thermal physics, it was less adventurous. Heat radiation is a thermal system. Einstein found in that system the same observable signature of discreteness as he found in ideal gases and dilute sugar solutions. So Einstein merely needed to hold true to the methods he had already developed to infer that heat radiation, under the conditions he specified, consists of many independent, spatially localized units of energy. What results is the light quantum.

The result follows from applying his statistical methods to heat radiation. Making sense of that result, however, proved harder and, over a hundred years later, the project remains incomplete.

Nineteenth Century...

Geometry

If the twentieth century was the century of novel physics, the nineteenth century was the century of novel mathematics.

Geometry had always been central to science, but it was languishing. Newton's mathematical techniques in his *Principia* of 1687 would have been immediately intelligible to Euclid himself.

That changed in the nineteenth century. There was an explosion of new ideas and methods. One of the foremost achievements of the century was a new conception of geometry. It included the idea of non-Euclidean geometries and their accommodation to yet more sophisticated geometries, notably projective geometry.

Felix Klein's "Erlangen Program" unified the many new geometries found in this century by means of another mathematical advance of that century, group theory.

Einstein...

General relativity

Einstein's general theory of relativity provided a qualitatively new way to think about gravity that is Einstein's signature novelty.

From a physical point of view, Einstein's theory was a bold departure. From a mathematical perspective, however, it simply applied nineteenth century mathematical techniques to a new and highly interesting application. Where Newton had employed Euclid's mathematics to display his account of gravity, Einstein employed the nineteenth century advances in geometry as the basis of his new theory.

The explicit framework was provided by Ricci and Levi-Civita's "absolute differential calculus," now called "tensor calculus." Einstein's mathematician friend, Marcel Grossmann, found Ricci and Levi-Civita's review article of 1901 on the calculus and drew it to Einstein's attention. It provided the framework he needed for his new theory.

2.3 Outlook

Nineteenth Century...

Unification

A major theme of nineteenth century physics was the theme of unification. The conception was that all the forces of nature were somehow related and that the burden of physics was to reveal those relations.

Nineteenth century physics is punctuated by successful unifications. Electromagnetic theory managed to bring electricity and magnetism together in the one theory. Light then proved to be merely a wave propagating in this electromagnetic field.

The single notion of energy unified many powers, such as heat, work and everything into which they may transform.

Ether

The grounding of nineteenth century electromagnetic theory was the ether. Electric and magnetic fields were not distinct processes, but were merely manifestations of different states of an all-pervading medium, the ether.

Einstein...

Unified field theory

Einstein's ambitions clearly held to this goal of unification. He had merged gravitation with the geometry of space and time in his general theory of relativity, completed in 1915.

In the decades that followed, he resolved to continue the unification. He now sought a single geometrized theory that embraced both gravity and electromagnetism, his unified field theory.

Einstein's metrical "ether"

Einstein famously did away with the ether; or, more precisely, he announced it superfluous and railed against the preferred state of rest attributed to it.

However, in his general theory of relativity and his unified field theories, Einstein retained an analogous background medium. It was not the ether of the nineteenth century. Rather it was a kind of geometrized version of it: the geometry of spacetime provided a substratum whose properties would be manifested as gravity and

Nineteenth Century...

Causation

The nineteenth century conception of causation was determinism: to say the world is causal is just to say that conditions now fix conditions in the future. This was a bare notion purged of the many finer aspects routinely assumed by a causal metaphysics.

Years later, Einstein himself described this nineteenth century conception:

"...the laws of the external world were also taken to be complete, in the following sense: If the state of the objects is completely given at a certain time, then their state at any other time is completely determined by the laws of nature. This is just what we mean when we speak of 'causality.' Such was approximately the framework of the physical thinking a hundred years ago."

Albert Einstein, "Physics, Philosophy, and Scientific Progress," International Congress of Surgeons, Cleveland, Ohio, 1950; printed in *Physics Today*, June 2005, pp.46-48.

electromagnetism. Indeed, as a concession to Lorentz, for a short time around 1920, Einstein talked of the metrical field, the carrier of geometrical properties, as an "ether."

Einstein...

Objections to quantum theory

Part of the original shock of quantum theory was the sense that its stochastic laws deprived the world of its causal character in this nineteenth century sense. There is a tendency now to discount Einstein's complaint against quantum theory, "God does not play dice." However it was repeated so often by him that we surely must take it as heart-felt. On its face, it is an honest expression of the nineteenth century alarm at the loss of causation.[1]

Einstein was quite nineteenth century in his expectation that the probabilities of quantum theory would somehow emerge from the supposed incompleteness of quantum description; that was precisely how the probabilities of statistical physics of the nineteenth century arose. Einstein's positive hope was that physics would continue along the lines of his general theory of relativity. Somewhere in his efforts to extend the theory to electromagnetism, Einstein hoped, the odd quantum phenomena would emerge. These hopes hold the quantum up to a nineteenth century ideal of a field theory in which notions of separability and locality are most fully implemented.

3. Einstein as the Bend in the Road

Imagine that we come to a bend in the road, to use a metaphor of Thomas Kuhn's.[2] When we stand at the corner, we see clearly the road that we have passed and also the road that is to come. The bend belongs to both parts. After we have passed the corner, all we see is the new road and the bend that started it. We no longer see earlier part it completed. Einstein is the bend in the road that joins the nineteenth and twentieth centuries of physics.

"...belongs equally to both ..., or it belongs to neither."

This image of Einstein as a transitional figure at the bend in the road seems to me to balance most effectively his connections with the times before and after him. The broader society, both popular and scientific, has found it harder to locate Einstein propertly. *Time* magazine did well when it declared him "Person of the Century" in their last issue of the 20th century of December 31, 1999. That seems fair. However Einstein has also become a prescient figure whose hunches somehow anticipate every modern fad and excitement. "Einstein was right" has become a popular slogan amongst scientists, especially when they want to suggest that Einstein would endorse their latest discovery or conjecture.

Notes

1. It is interesting to speculate on just how much Einstein was troubled by the loss of determinism in quantum theory. Was is as important as his concern that quantum theory was giving up on the notion of a local reality? I'm not sure that Einstein ever makes quite clear in his writings which was the more troubling failing. However we have an indirect report from Wolfgang Pauli, a physicist who knew Einstein well. The following is from a letter Pauli wrote to Max Born on March 31, 1954:

> "...Einstein does not consider the concept of 'determinism' to be as fundamental as it is frequently held to be (as he has told me emphatically many times), and he denied energetically that he had ever put up a postulate such as (your [citation to a letter of Born's]): 'the sequence of such conditions must also be objective and real, this is, automatic, machine-like, deterministic'. In the same way, he *disputes* that he uses as criterion for the admissibility of a theory the question: 'Is it rigorously deterministic?'
> Einstein's point of departure is 'realistic' rather than 'deterministic', which means that his philosophical prejudice is a different one..."

Max Born, ed., The Born-Einstein Letters. New York: Walker & Co., 1971.p. 221.

2. Here is Kuhn's writing about Copernicus:
“To ask whether his [Copernicus’] work is really ancient or modem is rather like asking whether the bend in an otherwise straight road belongs to the section of road that precedes the bend or to the portion that comes after it. From the bend both sections of the road are visible, and its continuity is apparent. But viewed from a point before the bend. the road seems to run straight to the bend and then to disappear; the bend seems the last point in a straight road. And viewed from a point in the next section, after the bend, the road appears to begin at the bend from which it runs straight on. The bend belongs equally to both sections, or it belongs to neither.”
T. S. Kuhn, *Copernican Revolution*, p. 182.

From Kantian schematism to the system of experience's invariants: the coordination of concepts and spatio-temporal objects in Cassirer's philosophy

Summary

This paper analyzes Cassirer's account of the coordination of concepts and spatio-temporal objects. We shall see that, in contradistinction to Kantian schematism, Cassirer maintains that this coordination is not achieved by means of a third element (the schema), which albeit intellectual is nevertheless also sensible. Rather, in Cassirer's view, the coordination will take place through a specification of the concepts that should be sought "within the domain of concepts itself."

Keywords

Cassirer, Kant, concept, schema, object, causality

Introduction

It has been argued recently that the key to solving many interpretation problems of the Kantian doctrine of schematism is to determine correctly the object that is to be subsumed under categories by means of schemata. More precisely, it has been maintained that this object is not merely an appearance taken as an object in general, but an appearance considered as a concrete, empirically given object. From this viewpoint, the proper task of schematism is to guarantee the coordination of concepts (in particular categories) and

individual spatio-temporal objects.[1] The aim of this paper is to analyze an alternative proposal to the Kantian solution of the coordination problem, which can be found in Ernst Cassirer's philosophy. We shall see that, in contradistinction to Kantian schematism, Cassirer maintains that the coordination of concepts and objects is not achieved by means of a third element (the schema), which albeit intellectual is nevertheless also sensible.[2] Rather, in Cassirer's view, the coordination will take place through a specification of the concepts that should be sought "within the domain of concepts itself."[3]

The thread that will enable us to reconstruct Cassirer's argumentation will be his interpretation of the causality principle. We shall see that this principle is the ultimate condition upon which the possibility of the coordination of concepts and spatio-temporal objects depends.

The structure of the paper will be the following. We shall begin by discussing the general framework of Cassirer's investigations on the causality principle. This is the framework provided by the transcendental method (§1). Then, we shall analyze the problems that Cassirer finds in the Kantian doctrine of causality, in particular regarding the schematism of that category, once the transcendental method is assumed (§2). Later, we shall discuss the main structure of the system of physical knowledge, for it is within this system that concepts and spatio-temporal objects get coordinated (§3). Finally, we shall investigate the

[1] Caimi (2010).

[2] "Obviously there must be some third thing, which is homogeneous on the one hand with the category, and on the other hand with the appearance, and which thus makes the application of the former to the latter possible. This mediating representation must be pure, that is, void of all empirical content, and yet at the same time, while it must in one respect be intellectual, it must in another be sensible. Such a representation is the transcendental schema." KrV, B 177.

[3] Cassirer (1956), p. 166. Our emphasis.

transcendental role that Cassirer assigns to the causality principle, and we shall show how this principle makes the coordination of concepts and objects in the system of physical knowledge first possible (§4).

§1 Cassirer and the transcendental method

The Neo-Kantianism of the Marburg school conceives itself as Kantian not regarding the content of its philosophy but rather regarding the form of its philosophizing. Neo-Kantians stress that they do not accept dogmatically any result of the Kantian doctrine. They just adopt the only real legacy of Kant: the philosophical method. This method is the transcendental one.[4]

According to the transcendental method, philosophy should take a certain factum as the starting point for the reflection and proceed to seek the possibility conditions of that fact. In the case of theoretical philosophy, the fact to be considered is experience. But experience is identified with physico-mathematical science. In this sense, Hermann Cohen maintains that experience is given in mathematics and in pure natural science,[5] and, more precisely, in Newtonian science.[6] The task of transcendental philosophy, as a theory of experience, will be then to determine the conditions of possibility of Newtonian science.[7]

Cassirer adopts this Cohenian conception of the transcendental method. But, while Kant took as a fact the science of his time, Cassirer applies this method to the new facta provided

[4] Natorp (1912), p. 194. However, it should be pointed out that the expression "transcendental method" is not to be found in Kant's texts. See: Baum (1980).

[5] Cohen (1877), pp. 24 – 25.

[6] Cohen (1910), p. 32.

[7] Cohen (1918), p. 93.

by the progress of physico-mathematical sciences, which include, in particular, non-Euclidean geometry and the relativity and quantum theories. In this way, Cassirer thinks that, starting with Kantian presuppositions, it is possible and even necessary to take the philosophical investigation beyond the stage reached by Kant himself. This progress, Cassirer remarks, is just the reaffirmation of the spirit of Kantian philosophy, since "the purpose of the Critique of Pure Reason was not to ground philosophical knowledge once for all in a fixed dogmatic system of concepts, but to open up for it the 'continuous development of a science'; in which there can be only relative, not absolute, stopping points."[8]

Cassirer discusses the role of the causality principle in the framework provided by such an analysis of the possibility conditions of the new scientific facts, according to the transcendental method. In the next section we shall consider the problems that Cassirer finds in the Kantian interpretation of that principle.

§2 Cassirer on Kantian causality

Cassirer maintains: "Of all the various explanations of the causal concept offered in the Critique of Pure Reason, perhaps the most precise and most satisfying is the one in which it is said that the concept represents nothing but a direction for the formulation of definite empirical concepts."[9] At this point, Cassirer quotes the following passage of the first critique: "That everything that happens has a cause cannot be inferred merely from the concept of happening in general; on the contrary, it is this fundamental proposition which

[8] Cassirer (1923), p. 355.

[9] Cassirer (1956), p. 127.

shows how in regard to that which happens we are in a position to obtain in experience any concept whatsoever that is really determinate."[10] The causality principle is a rule that indicates how we should conceive and form our concepts in order that they can fulfill their transcendental task: that of turning mere appearances into objective knowledge.[11] Thus, the causality principle is a principle about cognitions and not about things or events.[12] Even though in our everyday use of the principle we identify things as causes and effects, such a use is misleading if we are looking for a scientific foundation of causality. For what we call thing in our everyday experience is a complex of conditions that should be analyzed until the authentic scientific causal judgments are reached.[13] Such a characterization of thing as a complex of conditions expresses the core of Kant's doctrine, according to which the concepts of lawlikeness (Gesetzlichkeit) and objectivity are connected in a synthetic a priori judgment: only by means of a lawlike ordering can appearances be referred to an object of experience. In this sense, Cassirer indicates: "Objectivity or objective reality, is attained only because and insofar as there is conformity to law –not vice versa."[14] Therefore, we do not cognize objects, as if they (logically) preceded their laws, but rather by means of these laws we cognize objectively, as far as we establish certain limits and permanent connections in the uniform course of experience.[15]

[10] KrV, A301 = B357.

[11] Cassirer (1956), p. 19. Similarly, Kant declares: "They [the pure concepts of the understanding] serve as it were only to spell out appearances, so that they can be read as experience." Prol, AA IV 312.

[12] Cassirer, (1956), p. 65.

[13] Cassirer, (1956), p. 21.

[14] Cassirer, (1956), p. 132.

[15] Cassirer, (1956), p. 137; Cassirer (1923), p. 303.

However, Cassirer maintains, in the deduction of the causality principle carried out in the second analogy of experience, Kant wrongly directs his inquiry to empirical things and events, instead of exclusively focusing on empirical knowledge.[16] According to Cassirer, Kant rightly maintains the logical preeminence of the concept of law upon the concept of object, but the implications of such Copernican inversion are not fully assumed by the Kantian analysis of the causality principle. Here, Kant still struggles against representations of things and substances, as though a causal connection could be established by merely considering successive states of the same thing and determining the earlier as cause of the later.[17] Thus, following the example used by Kant, the objective series of a boat going down the river should not be established, for Cassirer, simply by determining the upstream state of the boat as the cause of its downstream state. Rather, the determination of the objective series requires considering the forces at issue and, more precisely, the physical laws (of gravitation, hydrodynamics and hydrostatics) that govern the movement of the boat. According to Cassirer, "these laws are the real components of the assumed causal connection."[18] In doing this, however, difficulties appear, since the exact formulation of those laws demands the symbolic language of physics, which differs significantly from the language of "things."[19] The determination of the transcendental role of causality as a condition of the possibility of scientific knowledge requires an analysis much more precise than the one carried out by Kant in the second analogy of experience. In particular, it is

[16] Cassirer (1956), p. 60.

[17] Cassirer (1956), pp. 21 - 22.

[18] Cassirer (1956), p. 22.

[19] Cassirer (1956), p. 22.

necessary to give up any consideration of "things and events" and the underlying concept of substance in order to focus on the functional form of experience. For this reason, Cassirer criticizes Kant and declares: "Kant did not follow to the end the road which he took in his solution of the Humean problem."[20]

Cassirer shares the Kantian position concerning the dependence of the concept of object on the concept of law, but he rejects the way in which Kant describes the proper role of the causality principle, for Kant then seems to invert the direction of the dependence. In fact, one might think that Kant initially assumes certain states of an object (like the upstream and downstream states of the boat) which then in a second move are represented in a causal relationship in order to determine their temporal order. However, this reconstruction of Kant's argument would not be correct. Rather, for Kant, it is precisely the causality principle that enables an objective determination according to the concept of law: the principle of causality determines the objective series of experience by means of a lawlike connection of successive appearances. In contradistinction to what Cassirer suggests, Kant's argument tries to show that only the representation of a law of my subjective perceptions enables their objective reference. Accordingly, the upstream and downstream states of the boat are represented as objective states only with the application of the principle of causality. Thus, Kant maintains the dependence of the concept of object on the concept of law in his analysis of causality.

However, according to Cassirer, Kant's position still has another shortcoming. The mere thought of lawlikeness, which for Cassirer expresses the core of the concept of causality,

[20] Cassirer (1956), p. 59.

leaves indeterminate how this conformity to law is to be empirically realized. Therefore, Cassirer says, Kant stresses that the category of causality should be specified in order to be useful and applied to experience.[21] This specification is achieved by means of the transcendental schematism. It is precisely this doctrine that explains how categories (in particular causality) may be applied to empirical phenomena. But, according to Cassirer, the validity of Kantian schematism is confined to the framework of Euclidean geometry and Newtonian mechanics. Therefore, Cassirer points out: "it is precisely these schemata which have lost their universal significance through the discovery of non-Euclidean geometry on the hand and the results of the special and general relativity theories on the other."[22] In the same sense, Cassirer maintains that the "crisis of causality" produced by quantum physics is not a crisis of the concept of cause, but rather a crisis of the way in which that concept is empirically applied through schemata. Thus, "such schematization has been definitely limited through the advent of the quantum theory. We can no longer combine causality with space-time description, let alone amalgamate the two in the manner of classical physics."[23]

Nevertheless, Cassirer indicates, Kant himself presents a version of the causality principle in which the latter remains free from the conditions imposed by schematism.[24] This is the A version of the principle, that reads: "Everything that happens, that is begins to be,

[21] Cassirer (1956), p. 166.

[22] Cassirer (1956), p. 166.

[23] Cassirer (1956), p. 166. For an overview on Cassirer's reception of the scientific progress of late 19th century and early 20th century, see: Plümacher (1996).

[24] Cassirer (1956), p. 162.

presupposes something upon which it follows according to a rule."[25] In Cassirer's opinion, this formulation solely demands the possibility of connecting through rules that which happens, without presupposing anything about those rules. Causality just implies the mere conformity of natural events to law. But, in the proof of the causality principle, Kant takes a further step by introducing time through the schema of cause and effect. Finally, Kant relates causality and continuity. The cause does not produce the effect instantaneously, but in a certain time interval $t_b - t_a$, such that a real magnitude b-a increases through all its intermediate degrees from its initial value a in t_a to its final value b in t_b.[26] However, quantum theory rejects this continuity requirement by accepting that certain magnitudes may only have discrete values and vary from an initial to a final value without adopting the intermediate ones. Therefore, the connection between causality and continuity, as Kant understands it, should be abandoned.

Given this criticism of Kantian schemata, one might well expect that Cassirer would search for new transcendental schemata that could perform the task that the Kantian ones, dependent on an earlier stage of science, are no longer able to carry out. However, Cassirer's proposal is much more radical. The new scientific facts, upon which a philosophical investigation carried out according to the transcendental method finally rests, demand a reinterpretation of the whole problem of Kantian schematism. In this sense, Cassirer maintains: "Transcendental logic can thus no longer be connected with or be dependent on transcendental aesthetics, as was the case in Kant's system. The demanded

[25] KrV, A 189.

[26] KrV, B 253 f.

specialization, indispensable for the empirical use of the causal concept, must now be looked for within the domain of concepts itself."[27]

In the next section we shall see that the conformity to law required by the principle of causality does not get specified by means of non-conceptual conditions. Rather, Cassirer puts forward a mere logical specification achieved in a system of invariants of experience.

§3 Physical knowledge as a system of invariants

Cassirer conceives physics as a system in which three types of statements are to be distinguished: the statements of measurement results, the statements of laws and the statements of principles.[28]

The statements of measurement results are the first step of the transition from the world of senses to the world of physics. This step is characterized by the conversion of sense data into determinations that may be subsumed under mathematical concepts. That which is perceived is represented in terms of measure and number and the immediate sense apprehension leaves its place to experimental observation.

From the point of view of the extension of knowledge, the statements of measurement results constitute a clear progress, because by means of measurement instruments it is possible to go beyond the contingent limits of our senses, as we do when we study, e.g., the

[27] Cassirer (1956), p. 166. Our emphasis. Nuzzo analyzes the modification of Cassirer's position on the relationship between logic and time from that maintained in Substanzsbegriff und Funktionsbegriff to the one of the Philosophie der symbolischen Formen. According to Nuzzo, in a theory of social sciences, a theory of invariants is not possible any more. Rather, logic becomes a kind of Hegelian phenomenology. See: Nuzzo (1996), pp. 76 – 77.

[28] Cassirer (1956), pp. 29 ff.

lunar surface with a telescope or blood cells with a microscope. However, this extension is not the key point at issue here. In parallel to an expansion of our world image, a concentration takes place too. The variety of sensible qualities leaves its place to a few fundamental determinations, from which the richness of the sensible data should be explained.

This concentration makes a crucial modification of our knowledge possible. Multiple perceptions just make up an aggregate: the sensible qualities of a perceived thing are merely juxtaposed. Color, smell, flavor and texture of an apple are independent from each other. Any combination of these qualities may contingently take place. Unlike a mere aggregate, the properties of a physical object, such as, e.g., and ideal gas, are organized into a system. Thus, the modification of one property entails the modification of the rest of them. In this way, temperature, pressure and volume of the gas are not independent properties, but rather their values are interconnected in a necessary manner.

Such relationships between the properties of a physical object are expressed by a different type of statement: the statements of laws. Whereas the statements of measurement results are characterized by an unavoidable reference to a "here and now," law statements have the logical form "if, then." Accordingly, a law statement cannot be taken as a mere summary of a number of statements of measurement results. Laws do not connect in a hypothetical manner individual magnitudes to which we may ascribe a spatio-temporal index. Rather, laws relate classes of magnitudes. For this reason, the statements of laws are not reached by means of an (always controvertible) inductive inference that, starting from many cases,

aims at their totality. To the contrary, in a law statement the "here and know" viewpoint is completely abandoned and the representation of a necessary connection is reached.

However, the transition from statements of measurement results to statements of laws is not the end stage in the process of physical knowledge. For just as the multiplicity of properties of a physical object acquires unity through laws, these laws are in turn unified by means of principles. Such unification is accordingly expressed by a third type of statement: the statements of principles. While the statements of measurement results are individual and the statements of laws are general, the statements of principles are universal. These statements do not refer to individual magnitudes or classes of magnitudes, but they connect different domains of physical knowledge, such as optics, mechanics or electrodynamics. The differentiation of these domains is thought of as relative to the higher principle, which therefore grounds the differentiation and, at the same time, unifies the domains.[29] Physics does not stop in front of the multiplicity of its laws, but seeks rules that enable the transition from one law to another. These rules are principles. Examples of them are Carnot's principle, the principle of energy conservation and the principle of least action, to which we shall immediately return.

The different types of statements are invariants of different order.[30] The statements of measurement results express values of physical magnitudes that do not depend on the subjectivity of the one carrying out the measurement. For example, in the same place and at the same moment, one observer may be warm, while a second one is cold. In each case, the

[29] Cassirer (1956), p. 44.

[30] On this issue, see: Ihmig (2001), pp. 81 ff. Ihmig develops in extenso Cassirer's theory of experience's invariants in: Ihmig (1995). See also Ihmig (1996).

perception has only subjective validity, varying from observer to observer. To the contrary, the statement of measurement result that expresses the room temperature remains invariant, since it is the same for all observers.

However, even though the temperature value is invariant in this sense, the statement of measurement result contains a spatio-temporal index: temperature has the value T in place x at time t. When knowledge progresses from statements of measurement results to law statements, this index disappears. A law statement does not include the particular temperature value in a certain place at a certain moment, as would be the case if a law were nothing more than a condensed expression of a collection of statements of measurement results. Rather, temperature is present in the law as a magnitude class that gets connected with other classes in a way that is invariant regarding the values of those magnitudes at different times and in different places. For example, the laws of Boyle, Mariotte and Gay Lussac connect temperature, pressure and volume of an ideal gas according to a rule independently of the absolute values of the spatio-temporal coordinates.[31]

But, as we have seen, the multiplicity of laws is to be distinguished from the rule that unifies them. In this case, the rule is contained in a statement of principle. Physics investigates how different laws, in particular those governing different areas of physics, are logically connected to each other. In doing this, the clue is not to be found in the different kinds of facts, but on the equations that express the structure of those areas. Cassirer puts forward the principle of least action as the paradigmatic example. In its application to particular cases, this principle was already known by Heron of Alexandria, who used it to

[31] Cassirer (1956), p. 42.

find the reflection laws of light, while Fermat also deduced the law of refraction by means of a more extended and deeper version of the principle. Leibniz made use of the principle in mechanics and Maupertuis even founded a proof of God's existence upon it. Euler gave a rigorous formulation and an exact physical meaning to the principle and Lagrange presented it in a complete and precise manner. Later, Helmholtz enunciated it as a universal physical principle, the validity of which went well beyond mechanics. The principle of least action, under the name of Hamilton's principle, is now a fundamental principle of modern physics, both of relativity and quantum theory.

The essential feature of the principle of least action is that it is not bound to any determinate content, since it is a variational principle. The principle establishes that certain magnitudes should have a minimal value, but these magnitudes can be quite diverse, e.g., the path covered by light (Heron), the required time for light to cover a path (Fermat), the product of velocity and path length (Maupertius), the mean value of potential energy (Euler) or the difference between kinetic and potential energy (Hamilton). The demand that such magnitudes acquire an extreme value determines the general form of the laws of diverse physical disciplines, providing in this way a heuristic rule for the search and discovery of such laws. These laws will be multiple, but the principle will remain invariant against them.

The statements of physics join together according to their invariance degree. The statements of measurement results are invariant against the subjectivity of the observer. Nevertheless, they vary against laws. These laws are in this respect invariant, but they are

not invariant against principles. Rather, principles are invariant against laws, and in being so they unify different branches of physics.

We can therefore see a reciprocal conditioning among statements that provides to the whole a systematic character. Neither are laws mere aggregates of measurement results nor are principles mere aggregates of laws. Physical knowledge does not originate from "an sich" elements, which may have sense and meaning independently of their relationship with others and that are accommodated in a kind of knowledge pyramid. Instead, we just find a functional coordination, in which all statements take part: the statements of "lower" type are entailed and presupposed by those of "higher" type. Thus, the right geometrical symbol of the system of physical knowledge would not be a pyramid, but rather a sphere, like the one that Parmenides uses to describe being.[32]

§4 Causality and the coordination of concepts and spatio-temporal objects

Even though the transition between statements of different type amounts to a qualitative jump, since it is in each case a modification of the kind of invariance that the statements express, the consideration of the causality principle entails a much more radical move. The causality principle is neither a metaphysical statement about the world in itself, nor an empirical statement about the sensible world, like those we have discussed so far. The causality principle does not talk about objects, but rather about our knowledge of objects and it is in this sense a transcendental principle.[33] More precisely, the principle is a statement about our empirical knowledge of objects and, thus, about the statements of

[32] Cassirer (1956), p. 35.

[33] Cassirer (1956), p. 58.

measurement results, about the statements of laws and about the statements of principles. According to Cassirer, the causality principle declares that all these statements "can be so related and combined with one another that from this combination there results a system of physical knowledge and not a mere aggregate of isolated observations."[34] In other words, the causality principle states that the conversion of sensible data into measurement results, their ordering according to laws and the unification of the multiplicity of these laws under principles is always possible. Thereby, even though such a process of systematization can never be considered as complete, its achievement should be sought as if an ultimate system were possible, by assuming that natural phenomena do not resist being systematically ordered.

Therefore, Cassirer understands the causality principle in a transcendental sense, as a condition of the possibility of scientific knowledge, but he ascribes to it a meaning that does not coincide with the Kantian one. According to Kant, the causality principle is constitutive for the possibility of experience, in so far as it makes the distinction between the subjective series of perceptions and the objective series of experience first possible. For Cassirer, the causality principle has instead a regulative character,[35] guiding our understanding towards the systematic unity of experience.[36]

[34] Cassirer (1956), p. 60.

[35] At this point, Cassirer agrees with Helmholtz, for whom the law of causality expresses that regulative principle of our thought that requires us to always look for more general laws: Cassirer (1956), pp. 61 ff. On the constitutive/regulative distinction in Cassirer's philosophy, see: Pätzold (1996).

[36] Cassirer indicates that, if we wanted to use Kantian terminology, we should call the causality principle a "postulate of empirical thought", since, as a purely methodological principle, it does not concern the content of the different types of statements but only the character of their objective validity. See: Cassirer (1956), p. 60. But, against Cassirer's interpretation of causality it should be pointed out that the causality principle in

Kant distinguishes the necessary task of three different cognitive faculties, which cooperate in knowledge: sensibility, understanding and reason. Sensibility provides us with intuitions. Understanding synthesizes these intuitions by means of concepts and thereby refers them to an object. Reason brings about the systematic unity of such objective cognitions. "Thus," Kant declares, "all human knowledge begins with intuitions, proceeds from thence to concepts, and ends with ideas."[37] From this viewpoint, the subsumption of spatio-temporal objects under concepts is the problem that the theory of schematism deals with. Schemata are precisely those representations that enable the spatio-temporal objects given by sensible intuitions to be thought by the concepts of understanding.

Cassirer, in contradistinction to Kant, does not pose the problem in terms of cognitive faculties, since in that way the danger of psychologism seems unavoidable.[38] The problem of the coordination between concepts and spatio-temporal objects is not that of the heterogeneity between intellectual concepts and sensible appearances.[39] Cassirer rather assumes a transcendental perspective from which there is just one single objectifying function. In this regard, Cassirer indicates: "According as the function of objectivity, which is unitary in its purpose and nature, is realized in different empirical material, there arise

modern physics does perform the very specific task of guaranteeing the objective character of the series of events, as it can be most clearly seen in relativity theory. On this issue, see: Schmitz-Rigal (2002), pp. 277 ff.

[37] KrV, B730.

[38] In fact, according to Cassirer, the true subject of the theory of schematism is the problem of the psychological possibility of a general concept. Cassirer (1922), p. 713. See also: Plümacher (1996), p. 119. However, Kant does not seem to aim at explaining here how a concept is formed, but rather how an already formed concept is applied.

[39] "Fassen wir den Verstand nicht lediglich als ein Vermögen der abstrakten Gattungsbegriffe, sondern, wie wir es nach der transzendentalen Deduktion der Kategorien tun müssen, als das „Vermögen der Regeln" auf, so hört er in der Tat auf, der Anschauung völlig „ungleichartig" zu sein." Cassirer (1922), p. 716.

different concepts of physical reality; yet these latter only represent different stages in the fulfillment of the same fundamental demand."[40] The transcendental task of each Kantian faculty is thereby reinterpreted as a different moment of fulfillment of that unitary function.[41] The statements of measurement results provide us with spatio-temporal data that are to be brought under rules expressed by statements of laws, the unity of which is attained by statements of principles.

However, since "in all scientific knowledge laws precede objects," Cassirer stresses that no object is 'given' to us "other than through laws."[42] Thus, the Kantian distinction between the sensible conditions under which objects are given in intuition and the intellectual conditions under which objects are thought by means of laws cannot be maintained any more. The data for objective knowledge, the "statements of the first level," cannot be isolated from statements of higher order as if "there would always be the possibility of imagining the higher layers removed without destroying the bottom layer or even altering it essentially."[43] To the contrary, "everything significantly factual is already theory."[44]

According to Cassirer, each type of statement expresses a peculiar moment of the conformity to law demanded by the causality principle. More precisely, each moment

[40] Cassirer (1956), pp. 137 – 138.

[41] "In diesem gesetzmäßigen Aufbau der Erkenntnis, in der Stufenfolge von Anschauung, Verstandesbegriff und Idee wir für uns alle empirische Wirklichkeit erst fassbar." Cassirer (1921), p. 61.

[42] Cassirer (1956), p. 143.

[43] Cassirer (1956), p. 35.

[44] Cassirer (1956), p. 35.

corresponds to a certain order of logical invariance. In this way, concepts and spatio-temporal objects get coordinated in a system of invariants of experience.

References

BAUM, Manfred. Article: "Methode, tranzendentale." In: Historisches Wörterbuch der Philosophie. Hrsg. v. Joachim Ritter und Karlfried Günder. Bd. 5, Basel – Stuttgart, 1980.

CAIMI, Mario (2010). "Der Gegenstand, der nach der Lehre vom Schematismus unter die Kategorien zu subsumieren ist." Conference held at the 11th Kant Congress, Pisa, Italy. Forthcoming in: Bacin, Stefano; Ferrarin, Alfredo; La Rocca, Claudio; Ruffing, Margit (eds.), Proceedings of the 11th International Kant Congress, de Gruyter.

CASSIRER, Ernst. Das Erkenntnisproblem in der Philosophie und Wissenschaft der neueren Zeit. Bd. 2. Berlin: Bruno Cassirer, 1922[3].

CASSIRER, Ernst. Substance and Function, and Einstein's Theory of Relativity (translation by William Curtis Swabey and Marie Collins Swabey). Chicago-London: Open Court, 1923.

CASSIRER; Ernst. "Goethe und die mathematische Physik". In: CASSIRER; Ernst. Idee und Gestalt. Berlin: Bruno Cassirer, 1921.

CASSIRER, Ernst. Determinism and Indeterminism in Modern Physics. Translation by O.Theodor Benfey. New Haven: Yale University Press, 1956.

COHEN, Hermann. Kants Begründung der Ethik. Berlin: Ferd. Dümmler, 1877.

COHEN, Hermann. Kants Begründung der Ethik. 2. Auflage. Berlin: Bruno Cassirer, 1910.

COHEN, Hermann. Kants Theorie der Erfahrung. 3. Auflage. Berlin: Bruno Cassirer, 1918.

IHMIG, Karl-Norbert. “Cassirers Rezeption des Erlanger Programms von Felix Klein.” In: PLÜMACHER and SCHÜRMANN (1996), pp. 141 – 163.

IHMIG, Karl-Norbert. Cassirers Invariantentheorie und seine Rezeption des 'Erlanger Programms'. Hamburg: Meiner, 1997.

IHMIG, Karl-Norbert. Grundzüge einer Philosophie der Wissenschaften bei Ernst Cassirer. Darmstadt: Wissenschaftliche Buchgesellschaft, 2001.

KANT, I. Critique of Pure Reason (KrV). Translation by Norman Kemp Smith. London: Macmillan, 1929.

KANT, I. Prolegomena to Any Future Metaphysics (Prol). Translated and edited by Gary Hatfield. Cambridge: Cambridge University Press.

NATORP, Paul. “Kant und die Marburger Schule”, Kant-Studien **17**: 193 – 221, 1912.

NUZZO, Angelica. “Das Verhältnis von Logik und Zeit bei Kant und Cassirer”. In: PLÜMACHER and SCHÜRMANN (1996), pp. 59 – 80.

PÄTZOLD, Detlev. “Cassirers Symbol-Formen: konstitutives oder regulatives Apriori der Repräsentation?” In: PLÜMACHER and SCHÜRMANN (1996), pp. 187 – 203.

PLÜMACHER, Martina. “Die Einheit der Regel der Veränderung. Zur Bedeutung der Wissensschaftsrezeption für Cassirers Theorie des Wissens.” In: PLÜMACHER and SCHÜRMANN (1996), pp. 113 – 140.

PLÜMACHER, Martina and SCHÜRMANN, Volker (Eds.). Einheit des Geistes: Probleme ihrer Grundlegung in der Philosophie Ernst Cassirers. Frankfurt am Main: Lang, 1996.

SCHMITZ-RIGAL, Christiane. Die Kunst offenen Wissens. Ernst Cassirers Epistemologie und Deutung der modernen Physik. Hamburg: Meiner, 2002.

Living Together in an Ecological Community

David E. Schrader

There is perhaps no area of ethical thinking that pushes us to examine the foundations of ethical thought more than environmental ethics. Should we think of the ethical demands placed upon our behavior in terms of the maximization of pleasure over pain? If so, should it be human pleasure and pain or the pleasure and pain of all sentient beings? Should we think of those demands in terms of the maximization of human happiness or of some other notion of human well-being? Should we think of those demands in terms of the promotion of certain types of virtue? Should we think of those demands in terms of rules governing some sort of moral community, perhaps a Kantian "kingdom of ends" or a Jamesian "Ethical Republic?" The practical question, of course, is how we are to live our lives. In particular, how are we to conduct ourselves when what is involved is our behavior as it affects the environment in which we and our children, grandchildren, and descendants well into the future? The philosophical question is what kind of analytical framework can be help us to think more clearly about how we are to live.

To address the philosophical question adequately it is important to keep clear focus on the range of practical problems that arise in our interaction with our environment. Suppose that we adopt an ethical framework according to which we judge our behavior on the balance of pleasure over pain that we produce. As Peter Singer has rightly noted in a large body of work, if pleasure and pain are the key moral criteria, it seems arbitrary to privilege human pleasure and pain over pleasure and pain in other forms of sentient life. At the same time, if we adopt a principle of determining our behavior so as to promote pleasure over pain in whatever forms of sentient life they may arise, we find some seriously counter-intuitive consequences. Suppose

that we find ourselves in the wilderness needing food, confronted with a choice of killing a common white-tailed deer or an endangered caribou. If our ethical principle is simply promoting the highest level of pleasure over pain it would seem that we could equally well kill the deer or the caribou. Either would likely experience roughly the same level of pain in its death. We find ourselves with an ethical principle that has no place for consideration of species membership. Such an ethical approach is unable to support the broadly shared view that preservation of species is a good.

A number of philosophers have attempted to frame environmental ethics in terms of the alleged intrinsic goodness of various natural objects. Quite apart from the inadequacy of most of the popular arguments for the position, it also fails to provide an analytic framework for addressing species problems. Perhaps worse yet, it would fail to provide any principled distinction between caribou and broccoli. This would seem to be the case with any approach that takes individual entities in the environment, whether human individuals or individuals of other sorts, as the starting point of ethical analysis without understanding those individuals as, in some sense, parts of a larger whole.

These considerations lead me to the view that the pressing ethical problems that arise in our interaction with the environment in which we live provide important support for understanding ethical agents centrally as parts of some sort of community of interrelated parts. For reasons that will become clear over the course of this paper, the kind of community in terms of which ethical life should be framed is best rooted in William James's "Ethical Republic" rather than Immanuel Kant's "Kingdom of Ends."

In current thinking about the environment, sustainability has become a very fashionable topic of conversation. We are, for example, presently in the middle of what UNESCO has

declared as the Decade of Education for Sustainable Development. I suspect that this declaration has generated responses in virtually every UN member nation. Among the many organizational responses in the United States has been the Disciplinary Associations Network for Sustainability, group of which the American Philosophical Association is a member. The APA has affiliated itself with the Network because we philosophers are surely among those whose discipline has an important contribution to make to the great discussions on how human beings can live together on this planet in ways that will facilitate not only our own well-being, but also the well-being of our descendants into the distant future.

While sustainability is in many respects a rather recent public concern, philosophers have for over two thousand years been concerned with the question of how human communities can be structured and can function in ways that will facilitate human well-being that will be sustainable from generation to generation. From Plato and Aristotle to thinkers of the present day, writers on social and political philosophy have been concerned with conditions that tend to undermine societies' capacity to endure from generation to generation. The central difference between this long-standing concern for sustainable communities and the contemporary interest in sustainability is that earlier discussions focused on the sustainability of the social and economic environments in which people lived, while contemporary discussion expands to include the physical environment in which people live as well.

The basic premise for most discussion of sustainability is that the capacity for growth in any system is not unlimited. Aristotle famously argues that the state will be most sustainable if it is not too large, either in population or territory (*Politics*, VII, 4-5). The historical record is clear that our particular judgments at any given time about the limits on growth are regularly mistaken. However the claim that the capacity for growth is greater than we might be able to foresee is

clearly far removed from the claim that the capacity for growth is unlimited. John Stuart Mill's discussion of "The Stationary State," in his *Principles of Political Economy*, is particularly apt on this point as it relates to economic growth.

> It must always have been seen, more or less distinctly, by political economists, that the increase of wealth is not boundless: that at the end of what they term the progressive state lies the stationary state, that all progress in wealth is but a postponement of this, and that each step in advance is an approach to it. We have now been led to recognize that this ultimate goal is at all times near enough to be fully in view; that we are always on the verge of it, and that if we have not reached it long ago, it is because the goal itself flies before us. The richest and most prosperous countries would very soon attain the stationary state, if no further improvements were made in the productive arts, and if there were a suspension of the overflow of capital from those countries into the uncultivated or ill-cultivated regions of the earth. (334)

Mill makes two central and correct points here. First, he notes that the limits of growth do expand with increases in technological development, although not unlimitedly so. Second, he notes that the growth of the richest countries expands as they use their capital to exploit the resources of less economically advanced regions of the earth. Mill also takes the position, almost alone among the great political economists of his time, that the stationary state, the condition in which the quest for growth has essentially come to an end, is a more desirable human condition than that in which growth dominates human aspiration.

> I cannot, therefore, regard the stationary state of capital and wealth with the unaffected aversion so generally manifested towards it by political economists of the old school. I am inclined to believe that it would be, on the whole, a very considerable

> improvement on our present condition. I confess I am not charmed with the ideal of life held out by those who think that the normal state of human beings is that of struggling to get on; that the trampling, crushing, elbowing, and treading on each other's heels, which form the existing type of social life, are the most desirable lot of human kind, or anything but the disagreeable symptoms of one of the phases of industrial progress. (336)

Mill's point is that the quest for perpetual growth brings with it a certain social instability, the "trampling, crushing, elbowing, and treading" that pits class against class and person against person. That social instability does not promote a sustainable human well-being.

As I noted above, present-day talk about sustainability is concerned not only with the sustainability of the social or economic environment, but also with the sustainability of the entire environment, including the physical environment. We have learned from Mill's contemporary, Charles Darwin, that humans are an interactive part of an embracing physical environment, just as all other living creatures are interactive parts of an embracing physical environment. Just as other species of living creatures can adapt to or fail to adapt to their environments, so humans can adapt to or fail to adapt to their physical environment. As Mill notes with respect to our social and economic environment, we have very substantial capacity to adapt through "improvements … in the productive arts," but in the end we too are subject to the inexorable demands placed on us by our total environment.

One of the pioneers in developing the study of ecology in the United States was Aldo Leopold, a forester by training. In a very brief section of Leopold's iconic *A Sand County Almanac*, Leopold advocates what he called a "Land Ethic." While the land ethic is not developed in any detail in Leopold's work, it rests on two foundations that provide material for

the development of a philosophical foundation for sustainability. First, Leopold claims that ethics "has its origin in the tendency of interdependent individuals or groups to evolve modes of co-operation" (202). While this is largely simple assertion on Leopold's part, the claim that ethics involve the evolution of modes of social cooperation presents an understanding of ethics that seems very much akin to the ethical views of pragmatists like William James. James' ethics replaces Kant's "kingdom of ends" with an "ethical republic" in which humans negotiate their needs and demands in a set of ever-developing equilibria.[1] On this basis Leopold puts forward the basic claim that "All ethics so far evolved rests upon a single premise: that the individual is a member of a community of interdependent parts" (203), a claim that echoes James' view that ethical terms only acquire any meaning in the context of a "moral universe," a context in which humans interact in ways that place mutual and sometimes conflicting demands upon one another (MP, 148-150). For the purposes of this paper, I will simply accept Leopold's first foundational claim, understood in roughly the manner in which I have elaborated James' ethical views elsewhere.

Leopold's second foundational claim moves squarely in the direction of grounding a contemporary understanding of sustainability. The "community of interdependent parts" that constitutes Leopold's community of moral concern is not simply a community of human agents. Rather it is a community that "include[s] soils, water, plants, and animals, or collectively: the land" (204). Leopold's extension of the moral community is, at one level, obviously correct. At the same time, it is, at another level, profoundly problematic.

The level at which Leopold's extension of the moral community is obviously correct, lies in a recognition that is borne out by the knowledge of evolutionary biology that we have gained

[1] See David E. Schrader, "Simonizing James: Taking Demand Seriously," *Transactions of the Charles S. Peirce Society*, Vol. XXXIV, No. 4 (Fall, 1998), pp. 1005-28.

since the work of Darwin. We humans clearly are parts of a system of "interdependent parts" that includes that includes "soils, water, plants, and animals." The basic biological mechanism of natural selection is a process of species developing in response to the various factors that their environments present. Thus species develop in response to the availability of various forms of nourishment, the variety of predators that threaten to eat them before they are able to reproduce, competition with other species that require the same sources of nutrition, characteristics that affect mate selection, etc. In short, the various parts of an ecological system are interdependent in a multitude of ways that leaves each part dependent upon all the others.

The level at which Leopold's extension is problematic, lies in how this collection of interdependent parts can constitute a "community." The problem becomes clear if we focus on the contrast between Kant's idea of a "kingdom of ends" and James's idea of an "ethical republic." Kant's "kingdom of ends" is a "systematic union of different rational beings under common laws" (100). Kant thinks that, to the extent that we are all rational, we will agree about the content of that "common law." James, correctly and by contrast, recognizes the frequent "falsity of our judgments, so far as they presume to decide in an absolute way on the value of other persons' conditions or ideals" (CB, 132). The necessary response to the "falsity," on James's view, is that the Kingdom of Ends must be replaced by an Ethical Republic in which the various members of the community are able to debate and challenge one another, working out a system of morality as an equilibrium in which both their consonant and competing demands may be most fully met. Central to the Ethical Republic is the ability of human agents to communicate with one another, to voice their various concerns and demands in a public forum. Accordingly, just as the Ethical Republic does not generate unanimous agreement, so it also does not generate eternal agreement either. The social equilibrium developed in the Ethical Republic must be an

ever-emerging equilibrium. This must be so because of the fact that, for each of us, our knowledge of the concrete needs and concerns of the other members of our community is never perfect and final. Moreover, it must also be so because of the fact that we will encounter different people over time, with different needs and concerns – people, needs and concerns of which we may well have been unaware at earlier points in time. The social equilibrium should always, however, be improving, becoming more encompassing, as we learn from our interaction with those others.

So the question that requires a satisfactory answer before we can accept Leopold's second foundational claim is "How can the ecological community, or what Leopold speaks of as the 'biotic community', the community that includes 'soils, water, plants, and animals' be a genuine moral community?" How can the interdependent parts of that collection of things communicate their needs relative to each other in ways that will make it possible for them to constitute, if not a moral republic, at least a moral polity of some functioning sort? The answer is not an easy one.

The first thing to note is that the claims of various environmental philosophers to the effect that natural objects have intrinsic value or that we should treat nature with respect or with empathy are simply not very helpful. Consider what it means to treat another with respect or empathy. I suspect that it means something like treating the other as I would like to be treated in similar circumstances. It is important to recognize that this cannot mean that I should treat the other as I would want to be treated if the other were like me. That would involve the deep arrogance of failing to recognize that the other is, in fact, distinct from me, that the other has a unique history, unique connections to the things with which it relates, and thus is a neighbor of mine, not a clone of me. Rather it must mean something more like that I should treat the other as

I would want to be treated if I were in his or her circumstances. Here, of course, we run up against James' important caution about the frequent "falsity of our judgments, so far as they presume to decide in an absolute way on the value of other persons' conditions or ideals." There is a real problem in my presumption to know what it is like to be in the other's circumstances. Given that our knowledge is a least largely based on experience, it is regularly much more particular in scope than we are inclined to think. To some extent it is accurate to say that I know what it is like to be human, but that claim must surely be understood in the context of the fact that my experience of being human has been gained through the experience of a sixty-four year old, well educated, reasonably affluent, male American. I simply do not know very much about what it's like to be a young Indonesian woman living in a small village. Much less, of course, do I know "What it's Like to be a Bat," or a dog, or a cow, or a coho salmon. Even worse, it may not even make sense to talk about my knowing what it's like to be a bacterium or a broccoli plant.

In the case of human communities we can see the central strength of James' idea of the ethical republic. While, as I have noted, I may not know much about what it's like to be a young Indonesian woman, we can at least construct contexts in which it is possible for young Indonesian women to talk about what their lives are like. Human history has shown that those in positions of power are often reluctant to listen to or to hear claims that threaten their power or privilege. Nevertheless, the idea of the ethical republic raises a framework that shows conditions for moral community, at least in principle.

While the problem of communication would seem to make it impossible for any kind of ecological community containing "soils, water, plants, and animals" to constitute a Jamesian ethical republic, I think it is still possible for us to make sense of an ecological community of a

somewhat morally less robust sort. In the last part of this talk I want to present a sequence of communities or putative communities, starting with the full-blooded moral community that James speaks of as "the ethical republic," and moving progressively through related moral communities in which communication is increasingly problematic.

An ethical republic is a community characterized by what many political philosophers speak of as "deliberative democracy." The principles governing the behavior of members of the ethical republic emerge as the members communicate with one another in richly varied ways. Central to the richness of communication in an ethical republic is the level of communication that humans achieve through their use of language. In an ethical republic we treat those around us as genuine ends, and not mere means, not because we have reflected on the desires of fully rational agents and reasoned to what principles such agents could will, but because we have listened to those other agents and have engaged them in deliberative processes to come to some mutual understanding of what we, as concrete and limitedly rational beings, in fact will.

There is surely some irony now in that I will reverse the procedure of Aristotle. Rather than moving from the household to the polis, I move from the ethical republic to the ethical household. My point, of course, is to move from a community of adult, functionally, albeit limitedly, rational and communicative human to a community of humans that includes infants and children. All of us who have had the experience of raising children have occasionally experienced the unhappy baby. The problem with the unhappy baby is that it is often difficult to determine what makes the baby unhappy. Is the baby tired, and hence in need of a good nap? Is the baby hungry? Does the baby have some sort of pain that might indicate some illness or is it the routine pain that comes with new teeth cutting through the skin inside the mouth? The baby is, of course, part of the family, part of a very important form of community. We presume that

the parents or caretakers of the baby are concerned for the welfare of the baby. They want to do what will ease the baby's distress. But in order to know what to do they must determine what is causing the baby's distress. All those of us who have experienced unhappy babies know that this situation is not easy. Yet we also know that it is not hopeless. The baby's inability to speak and tell us precisely what is bothering it makes it more difficult to determine what the baby's needs are, but it does not make it impossible. Essentially, in such situations, we experiment, with the experimentation informed by both our own experience and what we have learned from the experience of others. We try various options that we think will ease the baby's distress. Quite commonly we succeed in fairly short order, and hit upon the right solution. Sometimes we recognize that the baby's distress is caused by something serious, but we find ourselves unable to determine the cause. In such cases we frequently take the baby to be examined by a pediatrician whose special training carries with it an enhanced ability to determine the causes of the baby's distress. A note of caution is in order, however. Quite commonly we succeed, but not always. We have all heard of stories in which babies have died because those around it were not able to determine adequately and in time what it was that caused the baby's distress. Because the baby cannot tell us what is bothering it, the process of determining how to respond is more difficult, frequently more time consuming, and accordingly sometimes not successfully accomplished. It is, in short, distinctly fallible.

My point in moving to the household community is to dislodge the hyper-rationalization of human communities that has been a feature of so much philosophy at least from Kant to Rawls. No one seriously doubts that children and even infants are genuine members of our human communities. At the same time, no one can seriously doubt that the ability of children and infants to participate in rationally deliberative elaborations and negotiations of their needs

and interests is, in varying degrees, limited. The ability of humans to participate in such elaborations develops from being virtually non-existent at birth to being quite well developed at some point in adulthood. It is clearly not an "all or nothing" phenomenon. The course of the development varies somewhat from individual to individual, yet follows fairly general patters studied by developmental psychologists. For all that, they are members of our moral community from birth.

There is one other putative community that I want to consider before moving to the idea of an ecological community, the traditional farm. The family community is now expanded to include animals and plants. I don't know how many of you have any experience with farming. I spent most of the first twenty-two years of my life on the farm that is still my parents' home. We had horses, cattle, sometimes pigs and sheep, two dogs, and several cats. We raised corn, oats, and hay. Like babies, animals and plants lack language with which to explain the nature of their distress. Like babies, even more than very young babies, animals do engage in rather complex forms of behavior that frequently provide significant information about the nature of their distress. Plants present a more difficult case.

Certainly on a traditional farm there is a relationship of mutual dependence and interaction among the humans, the non-human animals, the plants, and even the soil on the farm. Each provides food for the others. Dogs provide help to the humans by herding cattle and chasing predators. Cats hunt mice, rats, and other pests. Like an unhappy baby, an animal experiencing distress has no language to communicate the problem. Yet like babies, and often in even more sophisticated ways, animals clearly give certain behavioral indications that convey information about the sources of their distress. Farmers who have substantial experience with animal behaviors can frequently gain useful information about the nature of distress. Again, as

with babies, there will be cases in which, even after some experimentation, the farmer either will not be able to determine the problem on the basis of animals' behavior or will determine that the behavior indicates a problem that lies beyond the farmer's ability to help mitigate the distress. In such cases the farmer will regularly seek expert advice, frequently in the form of a veterinarian.

The situation with plants is more difficult. First, plants can manifest various signs that they are in distress, but they lack the behavioral capacities of animals. Second, the resources from which the farmer can get expert advice in interpreting the signs of distress are not as powerful. This, however, does not mean that there are no such resources. While there are no "plant doctors," in the sense in which veterinarians are "animal doctors," most states in the United States have long provided what are called "County Agricultural Extension Agents." Virtually every state in the United States has a "land grant" university with a College of Agriculture and a department of something like Plant and Soil Sciences. One of the historical and present purposes of these land grant universities has been to study the sciences that enable farmers to better understand problems they may encounter in the raising of livestock and crops. The work done by researchers at the College of Agriculture is communicated to the broader community through a set of structures in which the County Agricultural Extension Agents serve as the central points of contact providing working farmers with access to the substantial bodies of horticultural and plant science research produced by the faculties of the land grant universities. Thus a farmer may learn that the yellow pallor of young corn may be caused by poor drainage of the farmer's field, or perhaps by an undesirable chemical composition of the soil.

This last possibility highlights the fact that one of the resources with which the farmer must work in the tending of plants is the soil. Farmers cannot be successful without attending to

the health of the soils on which they grow the various plants that they tend. There is an intimate relationship between plants and soils. One cannot simply plant a random plant in random soil and expect successful growth. Soil must be of a suitable type, contain suitable nutrients, be situated so as to receive suitable drainage, etc., in order to grow particular plant life. Accordingly, soil can fail to be "healthy" in a number of ways. The experienced farmer has a fairly good level of experience-based understanding of healthy soil and of various conditions that we might speak of as distressed soils. And again, the farmer has available expert resources on which to draw in determining the causes of soil distress that go beyond the competence of even the most experienced farmer.

The sequence of different kinds of communities or quasi-communities that I have just outlined provides a first step in addressing the question of how we can speak of an ecological community. The farmer's concern with the health of the "soil, water, plants, and animals" and the farmer's attempts to act on that concern provides an entre to understanding the "biotic" community of which Aldo Leopold speaks. Obviously the ecological community cannot be much like the robust "ethical republic" that constitutes the moral community of adult humans. At the same time, the family, surely a form of human community, cannot be a robust "ethical republic" either.

A robust moral community, be it an ethical republic or a kingdom of ends, requires members who can make generally reliable, albeit fallible and corrigible, judgments about their own needs and interests. The fundamental advantage of James' "ethical republic" over Kant's "kingdom of end" its more realistic acknowledgement of the fallibility of our judgments about the needs and interests of others and a more reliable framework within which our judgments about both our own and others needs and interests can be corrected. The moment we introduce

children, the reliability of whose judgments about their own needs and interests vary considerably over a course of cognitive development, and even more so infants, who lack the linguistic capacities to participate in the normal human processes of asserting and adjudicating needs and interests in community settings, we have been forced to adopt a more modest and nuanced model of moral community.

If we accept, as we surely must, that the collectivity of humans constitutes a genuine moral community, and we accept something like the account that I have given above about how human communities incorporate the needs and interests of infants and children into the process of framing collective judgments about how we are to live together in a way that pays due heed to the needs and interests of the full membership of the human community, we are well on the way to giving a plausible version of a "land ethic."

The two foundational claims on which Leopold rests his advocacy of a land ethic are 1) that ethics rests upon the premise that the individuals are members of a communities of interdependent parts, and 2) that we humans live in a community of interdependent parts that "include[s] soils, water, plants, and animals, or collectively: the land." I have, I believe, established that community, in the morally relevant sense, does not require either full rationality or full communicative ability. The move from the human community to a biotic community, then, requires that we recognize a complex system of mutually interdependent parts that includes "soils, water, plants, and animals," and that we be able to give an account of how we can frame generally reliable, albeit fallible and corrigible, judgments about the needs and interests of the non-human parts of such a system.

I have looked at the traditional farm as a kind of intermediary between the family and the putative biotic community. The farm cannot constitute robust ethical republics because of the limits constraining exchange of information and exchange of needs among its parts. The even-more severe limits constraining exchange of information and exchange of needs among the members of the biotic community likewise prevent its moral character from being robustly democratic. Yet our increasing knowledge of our larger physical environment has made it quite clear that humans do live on this planet within a system of interdependent parts that includes soils, water, plants, and animals. And humans do have resources through which we can learn a considerable amount about the well-being of that system of interdependent parts. As I have noted, those resources do not give us the depth or the richness of information that we gain through communication in the full-fledged ethical republic of human adults. Nevertheless the information is considerable.

The key, I think, to the parent's regular success in gaining information about the well-being of the unhappy child and the farmer's regular success in gaining information about the well-being of the soils, water, plants, and animals that comprise the farm community is interest, another notion to which James devoted considerable attention. Parents are drawn, I take it, by love to take a passionate interest in the well-being of their children, unhappy babies included. Farmers are drawn to a strong interest in the well-being of the soils, water, plants, and animals that make up the farm because they recognize the mutual interdependence of all of the parts that are required for farming to function. The first condition that is required for humans to understand themselves as members of an ecological or biotic community is likewise interest. One of the challenges of the contemporary world is urbanization. People living in cities are significantly removed from the natural processes that constitute our biotic environment. As a

result, it becomes easy for people to ignore the well-being of the interdependent parts of that system, and even their interdependence itself, in a way in which it would not be possible for a farmer to ignore the well-being of the soils, water, plants, and animals that interact on the farm every day. I take it that an important part of the motivation for UNESCO's declaration of a Decade of Education for Sustainable Development has been the recognition of the need to use the resources of education to promote such interest. Undoubtedly we must not be overly idealistic about the extent to which even the best environmental education can generate interest in the well-being of the biotic community. Nevertheless, it would seem that education is really the only resource we have in a world where large numbers of people lack significant experience of their interdependency with the other parts of the biotic community.

The second condition necessary for humans to be positively contributing members of an ecological community follows from the first. Both parents and farmers are strongly motivated to seek and usually heed the advice of people with expert knowledge concerning appropriate aspects of the well-being of babies, on the one hand, and soils, water, plants, and animals, on the other. An audience like this one knows well that universities train, in addition to philosophers, pediatricians, veterinarians, and horticulturalists, people who acquire expert knowledge on various aspects of the system of interdependent parts that constitutes our environment. Clearly in the United States, and I suspect most other places as well, there are demagogic politicians who tell people that popular or profitable ideas concerning our physical environment have nothing to learn from the work of environmental scientists. Part of our job, both as citizens and as teachers, is to help our communities fully realize that we ignore the results of environmental science at great peril, just as we ignore the advice of pediatricians and veterinarians only at the peril of our children and our farms.

I conclude then that living together in an ecological community is a challenging possibility, but a possibility nonetheless. It requires recognition that an ecological community must be something more modest than a full-blown moral community, a polity less robust than the ethical republic. The chief requirements of this polity are a genuine interest in the role of those parts that lack robust means of communication with us, and a willingness to acknowledge the superior, albeit always fallible and corrigible, knowledge of the scientists who claim some sort of expertise concerning the working of those parts. These conditions both present serious challenges. Neither can be met easily. Even in the best of circumstances, neither can be met adequately. Those facts, however, do not mean that we cannot make significant progress. Just as the ethical republic is an ongoing project, never getting it fully right, but hopefully making slow and steady improvement, so we have some reason to hope that living together in a biotic or environmental community may likewise be an ongoing project at which we may make slow but steady improvement. In both cases the cost of failure to make such improvement is seriously threatening to our ability to live decent human lives.

Works Cited

Aristotle, *Politics*, in Richard McKeon, ed., *The Basic Works of Aristotle*. New York: Random House, 1941.

James, William. "The Moral Philosopher and the Moral Life," in *The Will to Believe*. Cambridge, MA: Harvard University Press, 1979. Pp. 141– 162 (abbreviated MP).

_____. "On a Certain Blindness in Human Beings," in *Talks to Teachers on Psychology and to Students on Some of Life's Ideals*. Cambridge, MA: Harvard University Press, 1983. Pp. 132 – 149 (abbreviated CB).

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*, H. J. Paton, trans. New York: Harper and Row, 1964.

Leopold, Aldo. *A Sand County Almanac*. New York: Oxford University Press, 1949.

Mill, John Stuart. *Principles of Political Economy*, Vol. II. New York: D. Appleton and Company, 1894.

Schrader, David E. "Simonizing James: Taking Demand Seriously," *Transactions of the Charles S. Peirce Society*, Vol. XXXIV, No. 4 (Fall, 1998), pp. 1005-28.

The Pitt model of trans-disciplinary validity: challenges and prospects

Drozdstoj S. Stoyanov, MD, PhD
Assoc. Professor, Department of Psychiatry and Medical Psychology, Medical University of Plovdiv, Bulgaria

Abstract

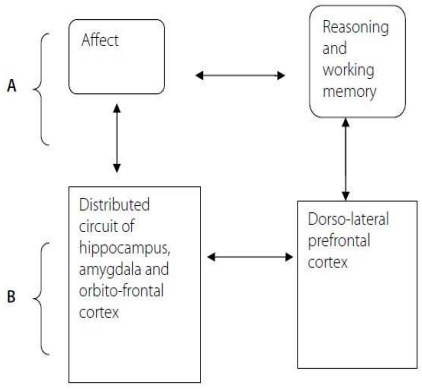
Evidence acquired inside the mono-disciplinary matrices of neurobiology, clinical psychology and psychopathology is deeply insufficient in terms of their validity, reliability and specificity, and can not reveal the explanatory mechanisms underlying mental disorders. Moreover no effective trans-disciplinary connections have been developed between them. In epistemological perspective current diagnostic tool are different but overlapping instruments exploring the same phenomenology. In line with the more scientific re-definition of mental disorders we defend the view that this process should take place under intensive (not extensive) dialogue with neuroscience. As Kato suggested only neurobiological studies using modern technology could form the basis for a new classification.This is to say that categorical approach in diagnosis should be abandoned in favor of broader diagnostic constructs (such as dimensional and prototype units) which are endorsed or "flanked" with data from neuroscience. Those broader units should be a subject of comprehensive evaluation of the personal narrative in context.

1. Epistemological foundations. Mind-brain translation

Many terms like “emotions” are frequently employed in many disciplinary systems, such as psychology, psychopathology, and neurosciences. Therefore “emotions” is a shared construct with all derivative terms which are used to describe human experiences in health and disease, e.g. “depression” or “anger”, “grief” and so forth. There are established vertical (bottom-up and top-down; reductive vs. emergent) connections in order to sustain biological explanations. My intuition is that there exists another kind of important ***horizontal*** **structures in mental health knowledge which require consistent translation**. We can define and understand the realm of mentality of its own right but any cognitive structure in it (regularity or notion representing certain aspects of consciousness) has to be underpinned with correspondent

cognitive structure in the realm of neural processes, which is demonstrated in the following diagram.

Diagram 1: Horizontal and vertical mechanisms



determining explanations in mental health knowledge

If regarded as *bridge psychophysical laws*, these lines of explanation must be supported with sufficient evidence to predispose reconciliation between paradigms in mental health and adequate translation between domains of neuroscience and humanities. However I do not embrace Nagel's demand for ontological elimination of the reduced entities besides the most basic sciences (neuro-biochemistry in our case). As it has been argued in a number of our previous papers a program for ultimate reduction (like those proposed by the eliminative materialism and eiphenomenalism) is predestinated to failure due to a number of meta-empirical reasons.

2. The impact of translation on the debate on psychiatric validity. The Pittsburgh model

Currently each of the disciplinary matrices (sources) concerned with mental health is discrete from the others. This means that neuroscience, clinical psychology and psychopathology employ operational disciplinary language and methods of its own right and has limited if any intelligible sense in the other two fields. In addition there is plethora of paradigm controversies which precludes the inter-disciplinary communication.

If we decide to focus for example on the case with the clinical assessment then a careful examination of the current methods reveals that a clinical psychiatric interview and a clinical

psychological rating scale consist of same kind of cognitive content. Nonetheless both psychiatrists and clinical psychologists claim that their tools (sources of information) are liable to mutual cross-validation.

Psychological tests (e.g. MMPI, Neo-FFI) are composed of self-evaluation reports (items) formulated as questions or statements. Psychopathological structured interview (e.g. PANSS, MADRS, DSM-IV structured interview) are typically formulated in the terms of subjective experience indicated as symptoms (actually these are self reports recorded by the physician) complemented with the so called 'signs' or the presumably 'objective' observations of the overt behavior of the patient.

The vast majority of the psychological assessment tools are standardized according to entirely "*atheoretic" empirical procedure*. In other words the items have been selected and keyed on the basis of their ability to distinguish diagnostic groups. The basis for the presumed "independent" assessment is actually the clinical judgment of the psychiatrist.

The current diagnostic hypothesis is raised and developed under the dominant psychiatric standard and is then is supplemented with the clinical psychological results. It is assumed that the psychological inventories are validated back to the psychopathological constructs and forward to the psychosocial outcome of the treatment.

Insofar none of the compounds of the structured psychopathological interview and the clinical psychological rating scales is independent to the inter-subjective patient and professional narratives. Both kinds of dimensions (psychiatric and psychological) lay inside the domain of *value-in* subjective assessment of human psychology, namely **the narrative**.

Therefore repeated protocols from various clinicians which serve to sustain the reliability claim of the 'scientific' DSM can not be regarded as independent measurement for the cognitive content and the value of the psychological rating scales or *vica versa*. Furthermore this undermines the potential to explain the mechanisms of production of mental disorders.

This is why they can not be credited as truly 'external' and 'independent' validity operations. We assert therefore that **only value-free facts from neurobiological research can play this role.**

At the same time the inter-disciplinary structure of psychiatry involves many facets from neuroscience which might regarded as one possible source of external validity as well as of explanation. Neuroscience shares many categories with psychopathology as mentioned above. However there are not introduced any relevant rules for "*translation*" of the data among these inter-connected domains of common interest.

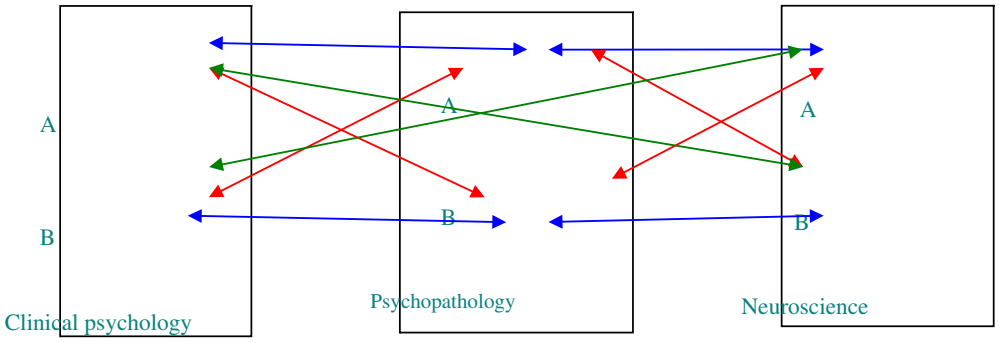
Instead in conventional context, expert committees have combined phenomenological criteria in variable ways into categories of mental disorders, repeatedly defined and redefined over the last half century. The diagnostic categories are termed “disorders” and yet, despite not being validated by biological criteria as most medical diseases are, are framed as medical diseases identified by medical diagnoses without any shared explanation of the mechanisms which produce them.

The more scientific re-definition of mental disorders requires intensive dialogue with neuroscience to reveal the connections with the person- and values-based understanding. One possible approach to achieve this goal would be to abandon the categorical taxonomy in favor of broader diagnostic constructs (such as dimensional and prototype units) which are endorsed or “**flanked**” with data from neuroscience. Those broader units should be a subject of comprehensive evaluation of the personal narrative in context.

The epistemic tool of this dialogue in my perspective should be the **translational cross-validation,** represented in the Pitt model bellow[1].

The figure bellow represents the trans-disciplinary nomothetic network of clinical psychology, psychiatry and neuroscience. They are regarded as three interconnected disciplinary matrixes, stabilized with cross-validity “bridging” structures. Each box illustrates a provisional common used term, where “A” stands for paranoia and “B” – for depression. The blue arrows indicate the bridges of convergent inter- and trans-disciplinary validity; the red arrows indicate divergent (discriminative) inter-disciplinary validity and the green arrows - discriminative trans-disciplinary validity.

[1] Known also as SMS paradigm, the latter has been coined most recently by C. Robert Cloninger from the names of its authors: Stoyanov, Machamer and Schaffner.



Experimentally this model entails simultaneous administration of clinical assessment tools with neurobiological measures. Its end-point implications are as follows:

(i) Replacement of the neutral (or inert) visual picture stimuli in current neuroimaging designs, since those can serve namely and only as ***neutral in terms of diagnosis emotional stimuli.*** This is to say that such kind of stimuli have no diagnostic value and hence can not relate or be embodied into validity operations and have no capacity to integrate explanatory mechanisms. To overcome this issue we suggest that the brain imaging should involve real-time rating with ***disorder-relevant scales*** (such as MMPI) performed simultaneously with fMRI.

(ii) Implement this approach to mental states and dynamic mental disorders e.g. bipolar depression;

(iii) Perform it as convergent validity operation between clinical and neurobiological measures in order to contribute to psychiatric validation and explanation.

3. Conclusion:

The basic claim of this paradigm is that the different sources (or disciplinary matrixes) of explanation should be enabled to converge toward an overlap of the neurobiological mechanisms underlying mental disorder and the narrative (s) of the patient in real time. Eventually such approach can facilitate the inter-domain translation of the shared terms in psychopathology, clinical psychology and neuroscience.

The introduction of such model has several consequences:

- It sustains the validity of the trans-disciplinary nomothetic network, which should exist to connect psychiatry, psychology and neuroscience
- Hence it sets a prerequisite for inter-domain translation
- Therefore it presents a prospect for bridging the explanatory gap between explanatory and 'understanding' domains of mental health knowledge.

ACKNOWLEDGEMENT:

Professors Peter K. Machamer and Kenneth F. Schaffner for their collaboration

REFERENCES:

1. Stoyanov, D, A linkage of mind and brain: towards translational validity between neurobiology and psychiatry, *Biomed Rev* 2011; 22: 65-76
2. Rivera-Hernandez, R, D. Stoyanov, Towards a pragmatic and functional unit of mind-and-brain. In reponse to Danko Georgiev's "A linkage of mind and brain: Sir John Eccles and modern dualistic interactionism", *Biomed Rev* 2011; 22: 85-89
3. Stoyanov, D., Machamer, P. K. and Schaffner, K. F. (2012), Rendering clinical psychology an evidence-based scientific discipline: a case study. *Journal of Evaluation in Clinical Practice*, 18: 149–154
4. Stoyanov, DS, Translational cross-validation among neuroscience and psychiatry: Prospects for diagnostic assessment and psychopharmacology, *European Psychiatry*, Vol. 27, Suppl. 1, 2012
5. Stoyanov, D., Machamer, P. K., Schaffner, K. F. and Rivera-Hernández, R. (2012), The challenge of psychiatric nosology and diagnosis. *Journal of Evaluation in Clinical Practice*, 18: 704–709
6. Stoyanov, D., Machamer, P. K., Schaffner, K. F. and Rivera-Hernández, R. (2012), The meta-language of psychiatry as cross-disciplinary effort: In response to Zachar (2012). *Journal of Evaluation in Clinical Practice*, 18: 710–72
7. Stoyanov, D, P Machamer & K. Schaffner, In Quest for Scientific Psychiatry: Towards Bridging the Explanatory Gap. *Philosophy, Psychiatry and Psychology*, in press
8. Stoyanov, D, P Machamer & K. Schaffner, A Fallacious forced choice: Cloninger and SMS are compatible, *Philosophy, Psychiatry and Psychology*, in press

The Missing Self in Hacking's Looping Effects. Forthcoming in *Mental Kinds and Natural Kinds*, H. Kincaid and J. A. Sullivan, eds. MIT Press.

The Missing Self in Hacking's Looping Effects

Part I. Introduction

Significant philosophical discourse has been dedicated to the ontological status of mental disorders.[1] The primary focus has been on whether mental disorders are natural kinds, i.e., whether they are similar to the kinds found in the non-human natural world, such as gold.[2] Ian Hacking argues that mental disorders are human kinds, differing from natural kinds insofar as they are subject to the looping effects of scientific classifications.[3] Mental disorders cannot be natural kinds precisely because being classified as having a mental disorder can bring on changes in the self-concept and behavior of individuals so classified. Such changes, in turn, can lead to revisions in the initial descriptions of mental disorders. Members of natural kinds, however, are not subject to such looping effects.

The phenomenon of looping effects is considered a compelling challenge to the claim that mental disorders are natural kinds and, as such, is discussed widely by both Hacking's followers and his critics. It is also widely resorted to by social scientists, especially those in critical disabilities studies, sociology and anthropology.[4] Yet the inherent complexity of the phenomenon has not been addressed, even by Hacking himself. In particular, the causal trajectory in which looping effects are generated and the way in which the subject responds to being classified remain unclear. Nor it is clearly understood how looping effects come about in the context of psychopathology. With a view to filling in some of these gaps, in this chapter, I note two connected shortcomings in Hacking's analysis of looping effects. First, his framework lacks an empirically and philosophically plausible account of the self to

[1] See for example, Hacking 1986, 1995a, 1995b, 2007a, 2007b; Cooper (2004a, 2004b, 2007; Samuels 2009; Graham 2010; Zachar 2001.

[2]There is no uncontroversial definition of natural kinds (Cooper 2004a, 2004b). Philosophers who discuss whether mental disorders are natural kinds mostly work with specific examples from the natural kind family, such as water, gold, animals etc. (e.g., Hacking 1986; Cooper 2004a, 2004b; Khalidi 2010). I follow their lead in this chapter.

[3]Feedback effects and looping effects are used synonymously by both Hacking and his critics. Throughout this paper, I use the latter.

[4] For some examples, see Carlson (2010); also Stets and Burke (2003).

substantiate the complex causal structure of looping effects. Second, he fails to engage with the complexity of mental disorder in the consideration of this phenomenon in the realm of psychopathology. Once the complexity of the selfhood and the complexity of the encounter with mental disorders are considered, it becomes clear that the causal trajectory of looping effects is more complex than hitherto envisioned.

Hacking uses the phenomenon of looping effects to articulate a dynamic nominalism, according to which the scientific classifications of human phenomena interact with those phenomena, leading to mutual changes. In other words, there is an interactive causal trajectory between scientific classifications and the subjects classified. Instead of describing what looping effects are, in reference to the features of the subject classified and the features of scientific classifications, Hacking uses examples to illustrate them. He includes not only mental disorders but also other human phenomena that are subject to scientific research, such as obesity, child abuse, refugee status. With these examples, Hacking shows how scientific classifications may generate changes in a subject's self-conceptions and behavior. However, a full discussion of looping effects requires *both* an account of the way in which scientific classifications influence the subjects *and* an account of *how* and *why* the subject responds to being classified in the way she does. Such scrutiny requires recognition of what the self is, how self-concepts are formed and how behavioral changes are motivated. In addition, when the phenomenon of looping effects is considered in the context of psychopathology, this scrutiny requires recognizing the complexity of the ways in which mental disorder influences the subject. The encounter with mental disorder changes self-concept and behavior, and it is not easy – if indeed possible – to discriminate the influence of diagnosis of mental disorder on self-concepts and behavior from that of the mental disorder itself. The fact that the diagnosed subject changes her self-concepts and behavior not only in response to being classified but also in response to her encounter with mental disorder reveals that the causal net of looping effects is much more complex than Hacking envisions. To the extent that he discusses the self

(he seems to be using self/person/subject/soul interchangeably),[5] he is informed by a simplified account of personhood, which situates the subject somewhere between genetic and neurobiological dispositions and freedom of choice. Hacking neither offers an account of mental disorders nor embraces the complex ways in which they shape people's self-concepts and behavior. Due to his superficial treatment of the self and mental disorder, he fails to make explicit the necessary and sufficient conditions for looping effects to be generated. This caveat makes his account the target of several partially successful criticisms.[6]

In this chapter, I offer a close reading of Hacking's work looping effects, evaluating his early and later works. Focusing primarily on the first arc of looping effects, i.e., how scientific classifications influence the subject classified, I show how he overlooks the complexities of the self and mental disorder. I then offer a model of the self, which I term the multitudinous self that substantiates the phenomenon of looping effects. To do so, in Part II, I expand on Hacking's work on looping effects and emphasize his dynamic nominalism – the key to understanding the features of looping effects. In Part III, I focus on his application of looping effects to mental disorders. In Part IV, I zoom in on Hacking's discussion of the self and indicate its superficiality. In Part V, I posit multitudinous self, a philosophically and empirically plausible model of the self that substantiates the complexity of looping effects in the context of psychopathology. This model of the self, I point out, can help scientific research programs to taxonomize mental disorders and can facilitate successful interventions into the lives of those with mental disorders, allowing them to flourish.[7] Thus, with the multitudinous self, I advocate a new style of reasoning about mental disorders in philosophy of psychiatry.

Part II. Dynamic Nominalism and Looping Effects

[5] For an example, see Hacking (2004).
[6] See for example, Cooper (2004a, 2004b); Khalidi (2010).
[7] Tekin (2011, 2010).

The phenomenon of looping effects is the linchpin of a series of works on what Hacking calls "making up people," which point to the way in which a new classification made by human sciences may bring a new kind of person into being.[8] Looping effects have a double arc. The first is the influence of classifications on those so classified, and the second is the ways in which some of those who are classified – and altered – modify the systems of classification. Some people with mental disorders (e.g. multiple personality and schizophrenia) are subject to the looping effects of psychiatric classifications; but looping effects are not restricted to the domain of mental disorders. Other examples Hacking uses include women refugees, pregnant teenagers, child abusers, the obese and, the genius.[9]

Hacking's dynamic nominalism is the metaphysical scaffolding for the phenomenon of looping effects; he explores "making-up people" by applying the realism versus nominalism debate to human phenomena.[10] The fundamental question in this debate is whether there is anything in reality that corresponds to universals, or whether there are only particulars. Realists accept universals into their ontology as mind-independent objects, i.e., they believe that universals are given by nature and exist independently of any perceiving human mind. Nominalists, on the other hand, argue that there are no universals, and they are not to be included in our ontology. All that exists are particulars, and it is human convention that individuates particulars, according to human interests. Hacking applies this query to what he labels human kinds, e.g., kinds of human beings, their embodiment, character, emotions etc.[11] He asks whether human kinds are given by nature, sorted and categorized independently of human intellect, or whether they are artifacts of human conventions. Does our naming, conceptualizing, and classifying individuate phenomena in the human world? Or are human kinds determined by nature prior to our ordering them? Hacking's traditional "static nominalist" would deny the existence of a mind-

[8] Hacking (1986, 1995a, 1995b, 1999, 2004, 2007a, 2007b).
[9] Hacking (1986, 1995a, 1995b; 2007a, 2007b).
[10] Although it is crucial to understanding the notion of looping effects, Hacking's critics have not discussed this metaphysical framework (e.g., Cooper 2004a, 2004b; Khalidi 2010)
[11] Hacking (1995b)

independent world sorted into neat categories,[12] holding that all classifications, taxonomies and classes are imposed by human conventions, not by nature. Over time, these categories become fixed. The traditional realist, in contrast, is committed to the idea of a naturally ordered world; as science progresses, we come to recognize and name pre-given categories. These categories are independent from humans; we discover them through science.[13]

Hacking's dynamic nominalism is situated somewhere between traditional realism and static nominalism. He believes that "many categories come from nature, not from the human mind."[14] However, these categories are not static, because the acts of sorting out, naming and classifying influence the individuals classified in those categories:

> The claim of dynamic nominalism is not that there was a kind of person who came increasingly to be recognized by bureaucrats or by students of human nature, but rather that a kind of person who came into being at the same time as the kind itself was being invented. In some cases, that is, our classifications and our classes conspire to emerge hand in hand, each egging the other on.[15]

Dynamic nominalism, situated as it is between traditional nominalism and realism, tracks interactions over time between the phenomena of the human world studied by the human sciences and the classifications of these phenomena. It is "realism in action," for Hacking, because "real classes of people" are sorted in new and specific ways; "making and moulding people as the events were enacted."[16] Another way of making sense of dynamic nominalism is thinking of it as "dialectical realism," as Hacking points out. Kinds of individuals come into being as an outcome of the dialectic between classifications and the classified. The naming of individuals as an outcome of scientific inquiry "has real effects on people," and such changes in people have "real effects on subsequent

[12] Hacking (1986, 1995b).
[13] Hacking (1986, 228).
[14] Hacking (1986, 228).
[15] Hacking (1986, 228).
[16] Hacking (2004, 280).

classifications." This phenomenon, for Hacking, can be captured neither by "an arid logical nominalism" nor by a "dogmatic realism."[17]

Hacking appeals to dynamic nominalism not only to elaborate on how sciences carve out human phenomena, but also to consider the implications of the study of human phenomena on the "possibilities of personhood."[18] Descriptions of human kinds influence the self-reflection of those human beings being described. Put otherwise, creating new ways of classifying people changes the subjects' epistemic and moral relations with themselves, including their self-concepts and self-worth. New ways of classifying even changes how these subjects remember their own past.[19] Hence, for Hacking, whenever philosophers think about persons as particulars, they "must reflect on this strange idea, of making up people."[20]

It is important to emphasize that even though dynamic nominalism provides the metaphysical scaffolding, there is no "uniform tale" or "general story to be told about making up people":[21]

> If we wish to present a partial framework in which to describe such events, we might think of two vectors. One is the vector of labeling from above, from a community of experts who create a "reality" that some people make their own. Different from this is the vector of autonomous behaviour of the person so labeled, which presses from below, creating a reality every expert must face.[22]

Although Hacking acknowledges the necessity to attend to both the scientific labeling from above, and individual's response from below, in making sense of looping effects, I argue his primary focus is on how human sciences influence and change the subjects they study. This is evident in his strategy to explain the phenomenon of looping effects: in accordance with his dynamic nominalism, he provides a plethora of examples to illustrate how human sciences generate changes in the individuals they study. However, as I show in Part III, his analysis of how the self – the subject of scientific study – responds to being classified remains superficial.

[17] Hacking (2004, 280).
[18] Hacking (1986, 230).
[19] Hacking (1995b, 369).
[20] Hacking (1986, 230).
[21] Hacking (1986, 233).
[22] Hacking (1986, 234).

Let me turn to Hacking's understanding of how human sciences induce changes in the subjects they study. The goal of these sciences is to acquire systematic, general, and accurate knowledge about puzzling and idiosyncratic phenomena pertaining to human beings in "industrialized bureaucracies," e.g., suicide, child abuse, multiple personality, obesity, refugee status. They seek to attain "generalizations sufficiently strong that they seem like laws about people, their actions, or their sentiments," so that helpful interventions can be made.[23] Unlike the objects of inquiry in natural sciences, the subjects of human sciences, i.e., human kinds, respond to how they are classified. Hacking demarcates between human and natural kinds by noting that human kinds are subject to looping effects due to the "self-awareness" of at least some of those classified:[24]

> Responses of people to attempts to be understood or altered are different from the responses of things. This trite fact is at the core of one difference between the natural and human sciences, and it works at the level of kinds. There is a looping or feedback effect involving the introduction to classifications of people. New sorting and theorizing induces changes in *self-conception* and in *behaviour* of the people classified. Those changes demand revisions of the classifications and theories, the causal connections, and the expectations. Kinds are modified, revised classifications are formed, and the classified change again, loop upon a loop.[25]

Hacking's best-known example of looping effects is multiple personality. Through this example, elaborated upon in the next section, the discussion of looping effects enters philosophical discussions of psychopathology, challenging the view that mental disorders are natural kinds.

Part III. Mental Disorders and Looping Effects

Hacking uses multiple personality as a "microcosm of thinking-and-talking about making-up people."[26] He wants to understand how "the sciences of the soul," in their attempts to make the soul an

[23] Hacking (1995b, 352).
[24] Two other traits distinguish human kinds from natural kinds. First, human kinds pertain to certain people and behaviors at a particular time, in a particular social setting, while natural kinds refer to the same kinds at all times. Second, human kinds are laden with social values, e.g., schizophrenia is a mental condition that is "bad" and is to be "healed." Natural kinds are value neutral, e.g., mud is not intrinsically good or bad (Hacking 1995b, 367).
[25] Hacking (1995b, 370, emphasis mine).
[26] Hacking (1995a, 5).

object of scientific query, make up people.[27] Thus, he is interested in the soul/subject/self/person[28] insofar as the soul is the *object* of scientific study; he does not consider the soul as a *subject*, i.e., he does not delve into what it is about the self that is prone to being made up.[29] This poses a problem concerning the details of the mechanism of the first arc of the looping effects, namely, what it is about the subject that makes her amenable to changing her self-concepts and behavior after being classified.

In Hacking's view, the popularity of the phenomenon of multiple personality among philosophers in the late 1980s and the 1990s stemmed from the challenges it posed to widely accepted conceptions of the self. Simply stated, it "refute[d] the dogmatic transcendental unity of apperception that made the self prior to all knowledge."[30] Hacking observes that the symptoms that characterize multiple personality disorder changed over time, as knowledge of the illness entered popular culture under the combined influence of curious psychiatrists, TV show producers, and alliances of patients. As Hacking sees it, those diagnosed with multiple personality start displaying different symptoms as they learn more about the illness and its manifestations in different individuals through popular culture, in a way that fits the popular descriptions of this condition. The changes in the symptoms they display, in turn, alter the classification of multiple personality. The following is a formulation of how looping effects are manifest in those with multiple personality:

> PM1: Psychiatry (as a human science) acquires systematic knowledge (K1) about human subjects (S1) who exhibit alternating personalities that are amnesic to one another. K1 picks out the perceived law-like regularities about S1 (e.g., alternating personalities).
>
> PM2: Based on K1, psychiatry forms classifications (CL1) of S1, labeling S1 "persons with multiple personality."
>
> PM3: At least some individuals with multiple personality become aware of their categories, as K1 is disseminated in popular culture through the combined impact of psychiatrists, TV show producers, alliances of S1a and so on.[31] (S1a), informed by K1, change their (b) *behavior and* (c) *self-concepts*.

[27] Hacking (1986, 1995a, 1995b).
[28] As noted above, he uses the self/soul/person/subject interchangeably, and I follow his lead.
[29] Tekin (2010, 2011).
[30] Hacking (1986, 224).
[31] Hacking (1999, 106).

> PM4: The awareness of being classified, the changes in the *behavior* and the changes in the *self-concept*s of those classified (S1a) amount to changes in the perceived regularities about these people. S1a, different from S1, starts to feature new symptoms; e.g., they exhibit animal personalities.
>
> PM5: Changes in the perceived regularities of S1a lead to changes in knowledge (K1) about their classifications (CL1), because S1a no longer fits the criteria for CL1.
>
> CM: Thus, classification of some people as "people with multiple personality" results in the creation of new knowledge (K1a), new classifications (CL1a) and new kinds of people (S1a) (e.g., according to K1a, people with multiple personality may exhibit animal personalities).

Hacking's claim that looping effects are not manifest in natural kinds is challenged by those who advance what I call the Parity Argument (PA), according to which there are looping effects in natural kinds comparable to those observed in human kinds, and the interaction between classifications and individuals is not exclusive to the human or social realm.[32] Proponents of PA suggest that our classificatory practices result in looping effects that alter some natural kinds, such as the influence of being classified as harmful on microbes, the influence of legal bans on the shape of marijuana, the influence of selective breeding on animals, and the influence of training on the domestication of dogs.[33] Corollary to PA is the failure of Hacking's claim that mental disorders are not natural kinds; if looping effects are not exclusive to human kinds but also are exhibited by natural kinds, it would be plausible to argue that those with mental disorders who exhibit looping effects can also be considered natural kinds.[34]

Hacking, in his early writings, apparently foreseeing such objections, attempts to clarify precisely what is unique about the looping effects in human kinds. He emphasizes, through different examples, that in the case of human kinds, because subjects are "aware" of "what we are doing to them," they are influenced by our "descriptions," and they change their self-concepts and behavior accordingly.[35] However, he is not consistent in his emphasis on the changes that occur in a subject after

[32] See Bogen (1988); Khalidi (2010); Cooper (2004a, 2004b).
[33] (Bogen 1988, Cooper 2004a, 2004b; Khalidi 2010; Douglas 1986).
[34] Cooper (2004a, 2004b)
[35] Hacking (1999, 106).

being classified. In particular, in some examples he postulates "being aware of being classified," "changes in self-concepts," and "changes in behavior" as individually sufficient changes that need to occur in the subject to generate looping effects (e.g., women refugees), while in others, all three are construed as jointly necessary changes for the looping effects (e.g., multiple personality). This inconsistency obscures his discussion of looping effects; it remains unclear whether these three variables are individually sufficient or jointly necessary for the looping effects to be generated. In his later writings, he adds new elements to the causal trajectory of the looping effects, but it remains unclear how and why the subject responds to being classified in the way she does.

In his early work, Hacking takes into account that the scientific classification of certain microbes as harmful and the resulting interventions influence these microbes. Such influence, however, is different from the influence of being classified on people:

> Elaborating on this difference between people and things: what camels, mountains, and microbes are doing does not depend on our words. What happens to tuberculosis bacilli depends on whether or not we poison them with BCG vaccine, but it does not depend on how we *describe* them. Of course we poison them with a certain vaccine in part because we describe them in certain ways, *but it is the vaccine that kills, not our words*. Human action is more closely linked to human description than bacterial action is.[36]

Hacking emphasizes here that in addition to the "intervention" facilitated by the classifications of human sciences, our "descriptions" guide subjects' self-directed feelings, concerns and actions, generating changes in their self-concepts and behavior. Natural kinds, on the other hand, are not subject to such looping effects: our words do not lead to changes in the self-interpretations of natural kinds; it is our interventions, *qua* classifications, that change them.

Elsewhere, Hacking develops this idea when he argues that naming and classifying, in and of themselves, do not make a difference in natural kinds: "the mere formation of the class, as separable in

[36] Hacking (1986, 230, emphasis mine).

the mind, and in language, our continuing use of the classification, our talk about it, our speculation using the classification, does not 'of itself' have the consequences."[37] To this, he adds:

> If N is a natural kind, and Z is N, it makes no direct difference to Z, if it is called N. It makes no direct difference to either mud or a mud puddle to call it 'mud.' It makes no direct difference to thyrotropin releasing hormone or to a bottle of TRH to call it TRH. Of course seeing that the Z is N, *we may do something to it* in order to melt it or mould it, or drown it, breed it, barter it...*But calling Z, N, or seeing that Z is N, does not, in itself make any difference to Z.* If H is a human kind and A is a person, then calling A H may make us *treat* A differently, just as calling Z N may make us do something to Z. We may reward or jail, instruct or abduct. But it also makes a difference to know that A is an H, precisely because there is so often a *moral connotation to a human kind.* Perhaps A does not want to be H! *Thinking of me as an H changes how I think of me*. Well, perhaps I could do things differently from now on. Not just to escape opprobrium (I have survived unscathed so far) but because I do not want to be that kind of person. *Even if it does not make a difference to A it makes a difference to how people feel about A – how they relate to A – so that A's social ambiance changes*.[38]

Note that in the above citations, Hacking emphasizes how the classification (or naming) changes the subject's epistemic and moral relations with herself. In other words, the category (the outcome of scientific query) into which the subject is placed, leads her to reflect on and judge herself differently. Being classified as A changes how she "thinks" about herself and her "self-worth." Such self-related epistemic and moral changes are generated through the scientific knowledge of the categories and are mediated *qua* self and *qua* others (who share the same cultural and linguistic community). Thus, in human kinds, naming and classifying *qua*-self and *qua*-others change the person. But natural kinds change only when naming and classifying lead to interventions.

Consider Hacking's response to a PA proponent, Mary Douglas.[39] Douglas, arguing for looping effects in microbes, suggests that microbes adapt themselves to the attempts to eradicate them (based on our classifying them as harmful) by mutating to resist antibacterial medications.

[37] Hacking (1992, 189-190).
[38] Hacking (1995b, 367-368, emphasis mine).
[39] Hacking (1986, 100-102).

This, in turn, eventually results in the modification of the classification scheme. To this argument, Hacking responds:

> My simple-minded reply is that microbes do not do all these things because, either individually or collectively, *they are aware of what we are doing to them*. The classification microbe is indifferent, not interactive."[40]

Hence, emphasizing the subjects' "awareness" of "what we are doing to them" and the change in their self-concepts and behavior is Hacking's way of distinguishing human kinds from natural kinds. However, he is not consistent in his emphasis on the "awareness" of being classified as a necessary condition for the generation of looping effects. Consider the following point about women refugees:

> A woman refugee may learn that she is a certain kind of person and act accordingly. Quarks do not learn that they are a certain kind of entity and act accordingly. But I don't want to overemphasize the awareness of an individual. Women refugees, who do not speak one word of English, may still, as part of a group, acquire the characteristics of women refugees precisely because they are so classified.[41]

Hacking presents women refugees' inability to speak English as a detriment to the degree they are "aware" of their labels and to the extent of the knowledge they acquire about their categorizations. Yet lack of awareness or limited access to knowledge about their labels does not prevent them from "acquiring the characteristics" associated with their category. How refugee women acquire these characteristics is not clearly articulated by Hacking, but it appears to be closely connected to their social cognition. A plausible explanation may go as follows. A refugee woman's interactions with others, who treat her as such, may lead her to change how she operates in the world and shape her behavior in a way that fits the label "women refugee."[42]

Proponents of PA, in developing the claim that natural kinds may be subject to the looping effects that Hacking attributes to human kinds, point out the ambiguity in Hacking's notion of

[40] Hacking (1999, 106).
[41] Hacking (1999, 32).
[42] Changes in behavior are explainable as the outcome of "socialization," a concept used in social psychology and sociology that is broadly defined as the way in which individuals are guided in becoming members of a social group. During their socialization, individuals conceptualize cultural knowledge like any other social information; they acquire, maintain, and apply these cognitive conceptualizations in their cognition and behavior (Kesebil, Uttal, Gardner 2010). The effects need not be conscious; indeed, they are often automatic. Women refugees may go through such socialization and unconsciously and automatically adapt to their labels.

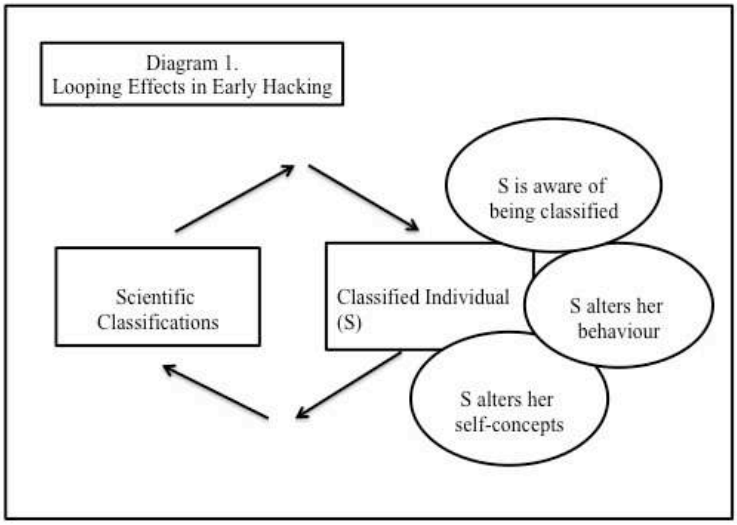
"awareness" and discuss whether it is a necessary condition for looping effects to be generated. For instance, Muhammad Ali Khalidi, a PA proponent, looks at Hacking's discussion of women refugee example.[43] For Khalidi, this example is a testament to the idea that awareness is not a necessary causal variable in the trajectory of looping effects. Thus, "awareness of being classified" does not demarcate human from natural kinds. Rachel Cooper, another PA proponent, also considers Hacking's emphasis on awareness. She suggests that awareness of being classified in itself does not show that human kinds cannot be natural kinds, because as it stands, Hacking's discussion merely shows that "human kinds are affected by a mechanism to which other kinds of entity are immune."[44] Although this indicates a difference between human kinds and other kinds, she does not take it to be fundamentally significant because "many other types of entity can be affected by mechanisms to which only entities of that type are vulnerable."[45] In other words, PA proponents conclude that awareness of being classified is not necessary for generating looping effects; thus, natural kinds can exhibit looping effects.

I will not develop it here in detail but in my view, PA proponents are seeking to deflate Hacking's emphasis on the subject's awareness of classification and the changes in her self-concepts upon being diagnosed. In particular, PA proponents neglect "the changes in self-concept" in Hacking's premises (PM3, PM4), taking the classification-induced changes in the subject to be primarily changes in behavior and interpreting these as culminating in "alterations in the kind." But Hacking himself fails to stress their importance: he does not offer a clear account of what a self-concept is, how self-concepts are formed or how exactly being labeled in a certain way changes a subject's self-concepts. In addition, as the PA proponents rightly point out, Hacking is ambiguous about whether awareness of being classified is a necessary variable in looping effects. While I agree with the claim that natural kinds are subject to some feedback effects, I contend that the types of causal loops exhibited in natural and human kinds, especially in the case of psychopathology, are significantly different from each other due to the

[43] Khalidi (2010).
[44] Cooper (2004a, 79).
[45] Cooper (2004a, 79).

complexity of selfhood and the complexity of the encounter with mental disorders. Once the shortcomings of Hacking's account are remedied by including an empirically and philosophically plausible model of the self to the trajectory of looping effects (see Part V), the types of differences between causal loops in natural kinds and those with psychopathology are explicit.

Diagram 1 summarizes the causal web of looping effects in early Hacking. Scientific classifications influence and alter the self-concepts and behavior of those classified; this, in turn, influences and alters the initial classifications.



My main concern with this framework is Hacking's reduction of the subject/soul/person/self to "classified person." Even when he considers the subject's awareness of her label and the alterations in

self-concept and behavior, he does not offer a detailed scrutiny of the self – the subject of classification. He does not explain what is involved in subjects' being "aware of what we are doing to them," or how people are influenced by "our descriptions of them" and change their self-concept and behavior accordingly. Is it a rudimentary level of awareness, or is it reflective and more elaborate? What motivates changes in self-concepts and behavior?[46]

In fact, Hacking's treatment of the "classified person" is superficial. This superficiality is problematic, especially when the phenomenon of looping effects is used in the context of psychopathology, as the subject of classification (or the clinical diagnosis) is also the subject of the mental disorder. In particular, the following three questions remain unanswered. First, how much of the change in the subject's self-concepts and behavior is connected to the knowledge she receives about the diagnosis? Second, how much is connected to the particular mental disorder to which she is subject? For instance, if the mental disorder is disruptive of her "awareness" and connected capacities for self-reflection, we need to take this into account. Aanosognosia in schizophrenia is a good example, as I note in due course. Third, how much of the change in the subject's self-concepts and behavior is connected to the clinical "treatment" she receives from mental health professionals upon diagnosis? It is hard to isolate these, as changes in the subject can be connected to a few, none, or all factors. Answers require a detailed scrutiny of the self and a close examination of the mental disorder. Although Hacking fails to consider these questions, they have important implications to understanding what looping effects actually are.

In his later work, Hacking partially responding to PA, advocates the abandonment of the notion of "natural kind" altogether and offers a framework within which to understand looping effects. In this latter discussion of looping effects, the causal net is wider; it includes, not only the classifications and the individuals classified, but also experts, institutions, and knowledge as key generators of looping effects.

[46] Some of these challenges are raised by PA proponents, as discussed above. See Khalidi (2010) for an overview.

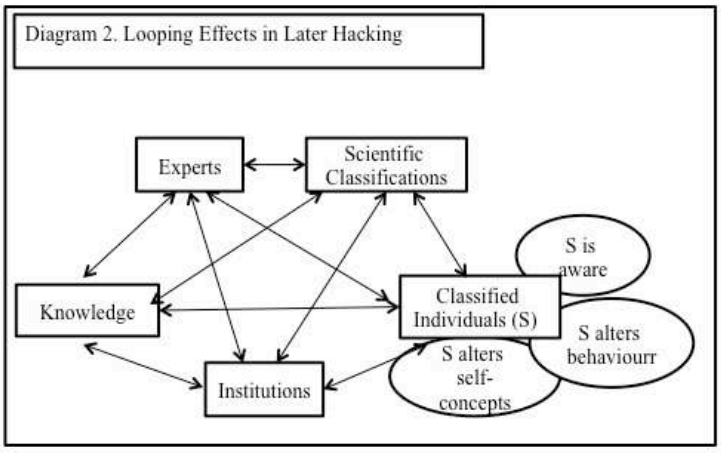
Consider first his abandonment of the concept of natural kind.[47] He argues that there are now so many radically incompatible theories of natural kinds that the concept has self-destructed. Some classifications, he suggests are "more natural than others, but there is no such thing as a natural kind."[48] This is not to say that there are not kinds in the world, but the idea of a well-defined class of natural kinds is obsolete.[49] The sheer heterogeneity of the paradigms for natural kinds, for Hacking, invites skepticism.[50] Calling something a natural kind no longer adds new knowledge; rather, it leads to confusion:

> Take any discussion that helps advance our understanding of nature or any science. Delete every mention of natural kinds. I conjecture that as a result the work will be simplified, clarified, and be a greater contribution to understanding or knowledge. Try it.[51]

Corollary to this change, Hacking no longer employs the term human kind when referring to human phenomena studied by the human sciences. Instead, he writes exclusively about the causal net of looping effects and instances of making up people, continuing to illustrate the phenomenon with examples.[52] He proposes a "framework for analysis" to understand the *kinds of people* studied by human sciences. In this new framework, the looping effects no longer occur on the two axes previously noted: *classifications made by human sciences* and *people so classified*. Rather, they occur between five axes, including the *experts* who classify, study, and help people classified, and the *institutions* within which the experts and their subjects interact. Additionally, there is an evolving body of *knowledge*[53] about the people in question, as well as *experts* who generate the knowledge and apply it in their practice. The interaction between these five elements leads to changes in individuals' self-concepts and behavior, as well as to changes in each component of this causal network, which, in turn, change the classifications.

[47] Hacking (2007a, 2007b).
[48] Hacking (2007b).
[49] Hacking (2007b, 205).
[50] Hacking (2007b, 207).
[51] Hacking (2007b, 229).
[52] Hacking (2007a).
[53] By "knowledge," he does not have in mind traditional epistemology's "justified true belief" but rather a Popperian conjectural knowledge.

Thus, while in his earlier writings, Hacking focuses on how classifications lead to the alterations in self-concept and behavior of persons, in the new and more complex framework, the other three elements are equally responsible. He points to the experts involved in the research on human phenomena, and the connected interventions, arguing that by their engagement in these activities, they influence the subjects they study. Similarly, the institutional framework within which these subjects are studied or helped also influence the subjects' self-concepts and behavior. Finally, the knowledge generated in this process is a mediator of change.[54] Thus, the causal net of looping effects, according to this new framework, is much wider. See Diagrams 2 below for an illustration.



Hacking's later framework is more responsive to how human sciences may generate changes in people's self-conception and behaviour, with the inclusion of the instruments though which these

[54] The influence of knowledge is discussed in early Hacking, but in his later work, he makes this influence more explicit.

changes are mediated. However, Hacking still does not explain what it is that about the individual that makes her respond to being studied in the way she does. Hacking continues to overlook the complexities of the "classified people" and the complexities of mental disorders they are subject to. The three questions raised above remain unanswered. It is still not explicit (i) how much of the changes in the subject's self-concepts and behavior are connected to the knowledge she receives about the diagnosis; (ii) how much of such changes are connected to her particular mental disorder; and (iii) how much of the changes in her self-concepts and behavior are connected to the clinical treatment she receives upon diagnosis. In other words, the course of illness and the influence of treatment remain excluded from the causal net of looping effects.

Let me illustrate with a paradigm case why these three questions are important. This case, depicting the complexity of looping effects, exemplifies why we need to know the complexities of selfhood and the complexities of mental disorders to understand how, why, and when looping effects occur.[55] Karl is a 26-year-old student working on a doctorate in music. He is known as a nice and respectful person. Although he is usually quiet, he opens up when he gets to know people. In his spare time, he teaches piano to children. He has two roommates. His dog has been with him since his early 20s. While studying for his PhD comprehensive exams, he starts to hear voices and see horrifying images. The voices are loud; they order him to do things he does not want to do, such as hitting the walls of his bedroom. He sees flames burning in his surroundings. He is unable to sleep. He talks to himself in an attempt to quiet the voices in his head. He is confused. Due to these orientational obstacles connected to his condition, he behaves differently at home and at school: he does not talk to his roommates and ignores the walking hours of his dog. Karl sees a specialist. After a few visits, the specialist decides that his symptoms are best individuated with the diagnosis of schizophrenia; she prescribes a kind of

[55] The example of Karl is informed by various memoirs of schizophrenia (e.g., Saks 2007) in a bid to show the complexity of mental disorder experience, something neglected by Hacking.

medication that, in her experience, is effective in reducing/completely removing hallucinatory symptoms.

Interrelated sets of conceptual and behavioral changes happen in Karl upon the diagnosis – the starting point of Hacking's looping effects. As Karl's illness unfolds, he continues to hear voices and talk to himself, but the visual hallucinations diminish with the help of the medication. The immediate changes are mediated by his illness; his treatment influences how he behaves and how he conceives himself. For instance, after noticing people's questioning looks when he is caught talking to himself, Karl spends less time in public spaces. For similar reasons, he stops giving piano lessons. His medication has side-effects, leading to more changes in his behavior: he sleeps too much and keeps his hands in his pockets to prevent them from shaking. The course of his illness and the treatment he receives also lead to alterations in his self-concepts. He used to consider himself a healthy person, fairly social, and a good dog-owner; now he considers himself ill and socially isolated, and contemplates giving his dog away as he is unable to care for him.

The knowledge he gains about his mental disorder, as well as the stereotypes associated with it, also motivates changes in his self-concepts and behavior. He surfs the internet, he consults books, and he reads the blogs and personal writings of other patients. He learns about aspects of his illness to which he was previously blind. After learning, for instance, that some schizophrenics have poor hygiene, he over-attends to his personal hygiene, to the extent that he annoys his roommates. Having encountered stereotypical representation of people with schizophrenia in the media about their inability to hold a job, he becomes skeptical of his ability to finish graduate school.[56] He considers leaving graduate school, fearing that he is not well- suited to becoming an academic. Yet at times, he wants to continue. He is confused.

[56] There is considerable evidence that stigma robs people with mental disorders of work, independent living, and important life opportunities (Corrigan, Edwards, Green, Diwan, and Penn 2001; Farina 1998; Farina and Felner 1973; Phelan, Link, Stueve and Pescosolido, 2000). Further, self-stigma may lead to impoverished self-esteem and self-efficacy (Corrigan and Holzman 2001; Corrigan and Lundin 2001; Wahl 1999).

Note that the changes Karl undergoes upon diagnosis are associated with (i) the knowledge he gains about his illness (including professional and cultural conceptions, as well as stereotypes), (ii) the course of his illness, and (iii) the clinical treatment he receives. Hacking's looping effects, applied to psychopathology, primarily targets (i). As cited above, Hacking suggests that changes occur in the subjects due to their awareness of being classified, and that "new sorting and theorizing induces changes in *self-conception* and in *behavior* of the people classified." In so suggesting, he takes knowledge about categories to be fundamental to the subject's changes. However, in the example above , the changes in Karl's self-concepts and behavior after diagnosis are not just mediated by (i), the knowledge Karl receives about his illness, but also by (ii), the course of his illness, and (iii), the psychiatric treatment he receives. It is difficult, therefore, to discriminate the influence of (i), (ii), and (iii) on Karl's self-concepts and behavior. If changes in the subject, i.e., "awareness of being classified," "changing self-concepts," and "changing behavior" are the fundamental generators of looping effects, Hacking must explain what precisely leads to these changes. The course of the mental disorder and the treatment the subject receives are as influential as her knowledge of the illness.

Nor does Hacking's addition of new elements to the complex causal structure of the looping effects in his more recent work answer these questions. While the explicit articulation of the interaction between institutions, experts, and knowledge, along with their separate and combined influence on the subject's self-concepts and behavior, shows that the causal net of looping effects is wider and more complex than originally envisioned, it remains unclear how and why the subject responds to these factors in the way she does. Hacking continues to consider the subject of human sciences as the "classified individual," and overlooks the complexity of the self that is subject to a mental disorder. To account for precisely how and why self-concepts and behavior may change upon diagnoses, he needs to take into account (ii), the course of the illness, and (iii), the subject's clinical treatment, not just (i), the knowledge the subject acquires about the illness.

These three questions can be answered by including the complexity of the self in the causal net of looping effects because the self is the *subject* of mental disorder, diagnosis, and treatment. The self is the agent of "awareness," as well as the agent of the changes in self-concept and behavior – the three causal variables of looping effects. It is also necessary to acknowledge the complexity of the subject's mental disorder. In Part V, I flesh out these contentions by including an empirically and philosophically plausible model of the self in its causal trajectory.

Part IV. The Self/Soul/Subject/Person in Hacking

Arguably, I am overstating my case, as Hacking did, in fact, albeit infrequently, write about selfhood. Be that as it may, my claim that Hacking's "classified individual" does not depict the complexity of selfhood is supported by his writing.[57] In "Between Michel Foucault and Erving Goffman: Between Discourse in the Abstract and Face-to Face Interaction" Hacking discusses his view of "making up people." Here, he clarifies his notion of "personhood," while developing his view that human sciences, in their classifications of people, their actions, and their sentiments, generate looping effects and make up new people. Hacking writes, and I cite at length:

> I must repeat my caution that there is not, and never will be any universally applicable theory of making up people. Just because dynamic nominalism is grounded in the intricacies of everyday and institutional life it will not lead to a general philosophical structure, system or theory. There is, nevertheless, a rather plausible general question in the offing. If we talk about making up people, we can sensibly be asked: 'What is your idea of a person, who can be thus made up?' I believe my own view was unwittingly formed in one of the heroic episodes of philosophy. Philosophy is heroic (in my version of events) when it tries to paint a picture of the *whole* human nature – and of the place of human beings in nature. Kant was heroic. Aquinas was heroic. Aristotle was heroic. I am the very opposite of heroic, not cowardly but proudly *particularist*. I think there is no fixed whole of human nature to discuss.[58]

This particularist stance is shaped by Sartrean existentialism. Hacking states that he relies on Jean Paul Sartre's conception of a person as a free individual with no essential features, who makes choices and creates his own destiny:

[57] Hacking (2004).
[58] Hacking (2004, 281, emphasis mine).

> We are born with a great many essential characteristics that we cannot change. Most of us can change how fat or thin, how trim or flabby our bodies are. But we can make only the most miniscule alterations to our height. A very great many physical characteristics appear to be fixed at the moment of conception, and many more are determined before the fetus sees the light. We do not yet have the genetic technology to change that, even if it were desirable. Neurologists and cognitive scientists teach us the same about the brain – that a great many of our potential thoughts and thought processes are innate, and that many more mental traits are part of our biological constitution. Many of the possibilities available to us, and many of the constraints imposed upon us, were dealt us at birth. At most we can choose what to do with what is there, although we know little except the most obvious facts about what is 'in our genes' and what is the result of other developmental processes. The chances of birth, of family, of war, of hunger, of social station, of the supports and the oppression that can result from religion or caste – the chances of wanton cruelty or high rates of unemployment – once you start listing everything there does not seem to be much room for choice at all. But of course there is. All that stuff is the framework within which we can decide who to be.[59]

It seems to me that Hacking places persons somewhere between "facticities" (to use Sartre's terms) – one's biological, genetic, neurological dispositions and limitations as well as social and cultural realities – and the "freedom" to choose whomever one wants to be in the face of these facticities, but he does not take into account the complexities involved in such placement. In other words, it is not straightforward to make choices in the face of facts; human decision-making capacities work in complex ways and do not allow one to "freely" make choices in the face of facticities. Consider, for instance, how he takes the existentialist motto "Existence precedes essence." Despite "constraints" to freedom, one can still choose:

> I favour an almost existentialist vision of the human condition over an essentialist one. But that vision is wholly consistent with good sense about what choices are open to us. We take for granted that each of us is precluded from a lot of choices for the most mundane of physiological or social reasons. Social: as a young man growing up in Vancouver, I could not have chosen to be an officer in the Soviet Navy. Physiological: my father thought I should spend my first two university years at a college that trains officers for the Royal Canadian Navy, because tuition was free, I would get free room and board, and it would make a man of me. Happily my vision was not good enough for me to be accepted. So I had the moral luck not to have to make a choice between a fight with my family and enrolling in the naval college.[60]

But while trying to avoid an essentialistic account of the self, an attitude consistent with his dynamic nominalism, Hacking stumbles upon a simplistic account of the self that is not responsive to

[59] Hacking (2004, 283).
[60] Hacking (2004, 286).

the complexities of real experience, the features of selfhood that make us responsive to our social and cultural environments and to scientific classifications.[61] This rather simplified account is not responsive to how selves actually experience the world, how they interact with others, how they develop self-related concerns and change their self-concepts, or what motivates behavioral change and how individuals make choices. Empirical evidence in cognitive sciences supports these intuitions about the complexity of human cognition. They offer perspectives on how the self interacts with the social world, how self-concepts are developed, what factors motivate behavior and behavioral changes, how the self experiences mental disorder, and how mental disorders shape behavior and self-concepts.[62] They point to the limitations of our computational capacities and those aspects of our reasoning processes that are driven by short-sighted reasoning strategies, cognitive biases, and opportunistic oversimplifications.[63] Such findings exhibit the complexity of selfhood and show that a Sartrean account is too simplistic. Most importantly, this simplistic account of the self does not enable us to answer the three questions raised above in the context of looping effects in psychopathology, i.e., how the subject's self-concepts and behavior change in response to (i) knowledge about the illness, (ii) course of the illness and (iii) the clinical treatment.

Part V. Multitudinous Self and Looping Effects

In what follows, I substantiate the complexity of looping effects in the context of psychopathology by including what I call the multitudinous self in its causal trajectory.[64] Multitudinous

[61] Feminist philosophers have criticized Hacking's neglect of the complexity of subjectivity and its inherent relationality, saying that, especially in his discussion of women's experience of multiple personality, he neglects the importance of oppression on the way women remember their past. In particular, Susan Campbell challenges Hacking's claim that the cultural acceptance of traumatic forgetting has allowed women to become suggestible to renarrating their past as having encountered and forgotten being abused as a child. Campbell criticizes Hacking's failure to consider social and relational influences on how women remember their past, and to politically analyze women's oppression (Campbell 2003, 192).
[62] Neisser 1988; Flanagan 1991; Nisbett and Wilson 1977. Pennebaker 1993. Miller, Potts, Fung, Hoogstra and Mintz 1990. Marin, Bohanek, and Fivush 2008; Jopling 2000.
[63] Gilbert 2006; Kosslyn 2006; Williams 2002.
[64] The inspiration for this model of the self is the poem "Song of Myself" by Walt Whitman, where he proclaims, "Do I contradict myself? Very well, then, I contradict myself; (I am large—I contain multitudes.)" Special thanks to Owen Flanagan who steered me in the direction of these lines, hence the word "multitudinous."

self is an empirically and philosophically plausible model of the self that captures the complexities of mental disorders and the process in which alterations occur in self-concepts and behavior. Multitudinous self is a dynamic, complex, relational, multi-aspectual, and more or less integrated configuration of capacities, processes, states, and traits which support a degree of agential capacity subject to various psychopathologies. To develop multitudinous self, I build on Ulric Neisser's account of the self as a complex configuration specified by various kinds of information originating from the subject and its social and physical environment.[65] Neisser argues that the forms of information that individuate the self are so different from one another that it is plausible to suggest that each establishes a different "self." Therefore, he distinguishes five separate selves: the ecological self, or the self who perceives and who is situated in the physical world; the interpersonal self, or the self embedded in the social world who develops through intersubjectivity; the extended self, or the self in time grounded on memory and anticipation; the private self, or the self exposed to private experiences not available to others; and the conceptual self, or the self that represents the self to the self by drawing on the properties of the self and the social and cultural context to which she belongs. All five selves are empirically traced by research in cognitive sciences, including developmental psychology, social psychology, cognitive psychology, and neuroscience. [66]

Instead of construing these five as distinct selves, I take them to be five aspects of the self, forming the multitudinous self. Each aspect is identifiable from the first and third person point of view. These aspects are instrumental in connecting the subject to herself and to the physical and social environment in which she is situated.[67] Multitudinous self can be construed as a self-organizing system of these five aspects, a locus of agency that remains more or less integrated through time. The ecological and intersubjective aspects of the self are based on perception and action and are present at the earliest

[65] Neisser, 1988.
[66] Neisser investigates each of these selves by appealing to a wide range of research in developmental, social and cognitive psychology. He edited and co-edited several volumes on the different selves. For example see Neisser (1993); Neisser and Fivush (1994); Neisser and Jopling (1997).
[67] Neisser (1988).

stages of human development. Meanwhile, the temporally extended, private, and conceptual aspects of the self are often grounded upon memory, reasoning capacities, the development of representational skills and language; they develop as the cognitive mechanisms mature.[68] The ecological aspect is grounded in the body and is specified by the physical conditions of a particular environment and the active perceptual exploration of these conditions by the subject.[69] It is continuous over time and across varying physical and social conditions.[70] The intersubjective aspect is individuated by "species-specific signals of emotional rapport and communication" between the self and others.[71] It appears from earliest infancy, as the infant engages in social exchange through interaction with caregivers.[72] The temporally extended layer of the self is grounded on what the self remembers and anticipates. It relies on autobiographical memory or other stored information.[73] What the subject recalls depends on what she now believes, as well as what she once stored. The private aspect of the multitudinous self contains the subject's felt experiences that are not phenomenologically available to anyone else (such as pain); it appears when children first notice that some of their experiences are unique to them.[74]

What is most important for the purposes of this chapter, is the conceptual aspect of the multitudinous self, because it hosts self-concepts, which are influential in guiding behavior. Self-concepts selectively represent the self to the self. They are the products of the dynamic interaction between the aspects of the self, and the features of the social and cultural environment. In turn, self-concepts inform and shape the aspects of the self as well as some features of the social and cultural

[68] See Neisser (1988); Jopling (1997); Pickering (1999); Gibson (1993).
[69] Neisser 1988). Eleanor Gibson calls this the "rock-bottom self" that collects information about the world and interacts with it (1993, 41).
[70] Jopling (1997, 2000).
[71] Neisser (1988, 387).
[72] See Trevarthen (1980); Neisser (1988); Fogel (1993); Murray and Trevarthan (1985); Bowlby (1969); Stern (1993).
[73] Bartlett (1932).
[74] It is difficult to determine when introspective reference to private experiences develops, but many studies show that children are aware of the privacy of their mental life before the age of five. The four-year-olds tested by Moessler et al., for example, clearly understood the notion of a "secret" (Moessler et al, 1976).

environment. Self-concepts are thus informed by the features of the four aspects of the multitudinous self, the subject's embodied experiences in the world (such as illness).[75] Let me consider them in turn.

Self-concepts include ideas about our physical bodies (ecological aspect), interpersonal experiences (intersubjective aspect), the kinds of things we have done in the past and are likely to do in the future (temporally extended aspect), and the quality and meaning of our thoughts and feelings (private aspect).[76] For instance, my self-concept as a "friendly person" is a product of the intersubjective aspect of my selfhood and of the norms of friendliness in the culture I am a part of.

Self-concepts are informed by the pathologies the person is subject to. This influence is mediated by: the changes that occur in the ecological, intersubjective, temporally extended, and private aspects of the self due to illness; the scientifically based or folk psychological knowledge available to the person about her illness; and her self-narratives in making sense of her condition.[77] For example, having lung cancer affects my ecological self by, say, making it difficult for me to breathe, and this may lead to alterations in how I conceive myself and limit my actions (I may decide to stop running outside). This, in turn, affects my self-concept about my body, something tied to my ecological layer (I may form a self-concept as a person who has difficulty breathing). Or consider Karl. Due to the voices he hears, he talks to himself. In order to avoid being seen speaking to himself, he stops taking public transit. His self-concept as a responsible person caring for the environment by using public transit may shift, in the light of his altered behaviour.[78]

Self-concepts are shaped by folk and scientific knowledge available to the subject about her illness. For instance, what Karl learns about the course of his illness from various scientific and folk media may lead him to alter his self-concepts. Prior to his illness, he considers himself someone who

[75] Neisser (1988).Jopling 1997, Tekin 2011
[76] See Jopling (1997, 2000); Neisser (1988).
[77] Tekin (2010, 2011).
[78] Of course, not every illness experience leads to alterations in self-concepts. People with psychotic disorders such as delusional disorder (once known as paranoia), and schizophrenia commonly suffer from anosognosia - that is, a lack of awareness of their disorder, its symptoms, and its severity (Amador, Seckinger 1997; Amador, Strauss, Yale and Gorman, 1991). Such psychiatric patients may not change their self-concepts in response to the illness experience.

wants to pursue a career in academia, but upon learning the scientific accounts of the course of his illness, he revises his self-concepts. In addition, the narratives Karl tells himself about his illness may alter his self-concepts.

Self-concepts are action-guiding; our ideas about ourselves inform how we behave. My self-concept of my physical strength affects my physical activities: I may or may not reach out to lift a suitcase depending on how strong I feel and how heavy I perceive the suitcase to be. Similarly, my self-concept about my intelligence and ability to learn new philosophical material influences what I can actually learn or how well I do in a job interview. Similarly, in the context of mental disorders, the self-concepts formed or altered in this vein influence subject's actions. For instance, Karl's concept of himself as a person with schizophrenia who will be unable to finish the graduate school may in fact influence his decision to quit the graduate program he is enrolled. Similarly, his self-concepts may constrain or expand his resources in responding to his illness.[79] Perceiving himself as someone who needs help he may reach out to communities of individuals who experience a similar condition. Thus, self-concepts motivate the subject to think, act, and behave in certain ways, restricting or expanding her possibilities for action.[80]

Note that the multitudinous self incorporates psychopathology in its structure, taking it as a possible feature of the self. Mental disorder is broadly construed in this model of the self by considering how well the subject functions with respect to the layers that connect her to her self, her social world, and the physical world; it takes the complexity of selfhood as the norm. As multitudinous self embraces the complexity of being subject to psychopathology, we can use it to make sense of how self-concepts change after the subject receives a diagnosis of mental disorder. Self-concepts and behavior change due to: (i) the subject's knowledge of the illness, as Hacking emphasizes in his discussion of looping effects; as well as (ii) the course of illness; and (iii) the psychiatric treatment the subject receives.

[79] Tekin (2010, 2011).
[80] Tekin (2010, 2011); Jopling (1997); Tekin forthcoming.

The multitudinous self illuminates the case study cited above. Karl's experience with schizophrenia can be traced through the five aspects of the multitudinous self. The symptoms of schizophrenia, such as hearing voices and encountering hallucinations, are part of the private aspect of the self. These can also be traced through the ecological aspect, insofar as some neurochemical changes are associated with such experiences. Schizophrenia compromises Karl's interpersonal relationships; he does not talk to his roommates and ignores the walking hours of his dog, phenomena linked to the intersubjective aspect of his selfhood. Schizophrenia may also compromise Karl's plans for the future and his feelings about the past, thereby affecting the temporally extended aspect.

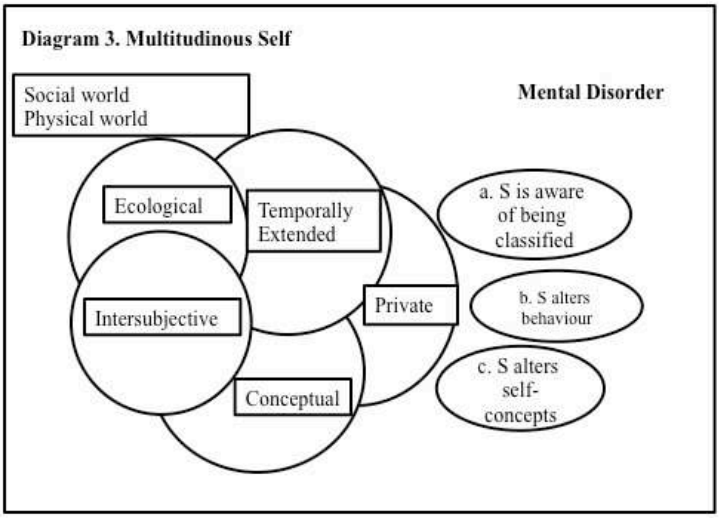
All these alterations in the way Karl experiences himself and the world change how he conceives himself and how he behaves. The diagnosis he receives, the psychiatric treatment that accompanies the diagnosis, the onset of schizophrenia, the social treatment he receives from his community, and the knowledge he acquires about his illness lead to interrelated changes in his self-concepts and behavior. As discussed above, some symptoms may diminish while others remain: although he may continue to hear voices and talk to himself, the visual hallucinations may diminish with the help of the medication. But other experiences may present themselves; he may start sleeping excessively, for instance, or he may become more socially isolated. His former conception of himself as a healthy person, fairly social, and a good dog-owner may be replaced by the idea that he is ill and socially isolated. Knowledge he gains about his schizophrenia, the cultural stereotypes and prejudices associated with it, and the self-narratives he creates will all influence his self-concepts and behavior.

Thus, the changes in Karl stem not only from (i) the knowledge he gains about his illness (including professional and cultural conceptions as well as stereotypes), as Hacking emphasizes, but also from (ii) the illness itself and (iii) the clinical treatment he receives. Thus, Hacking's discussion of looping effects, insofar as it emphasizes (i), is only the tip of the iceberg; the changes in those receiving a psychiatric diagnosis are more complex, given the dynamic and multilayered nature of selfhood and the complexity of the encounter with mental disorder.

Multitudinous self bolsters our understanding of looping effects by explaining how and why the self responds to being studied in the way it does. To sum up, three features of the multitudinous self framework permit such scrutiny: (i) multitudinous self explains the reflective influence of psychiatric diagnosis on people; (ii) it considers the illness experience as a part of the self-experience of the subject; and (iii) it explains how the clinical and intersubjective treatment the subject receives changes her self-concepts and behavior.

In short, multitudinous self is an empirically and philosophically plausible model of the self; the aspects of the self are responsive to experiences of actual people as we encounter them in daily life and can be scrutinized by multiple, interdisciplinary scientific analyses.[81] As unexplainable phenomena will remain despite the multiple approaches offered by various sciences and first-person accounts of selfhood, it is important to work with a model of the "self" rather than the particular layers of the self which can be clustered as, say, "genetic make-up," or "moral luck" (as Hacking does). Doing so prevents the reduction of a complex set of questions pertaining to the self and mental disorders. Without the multitudinous model of the self, in other words, we will lose important information about actual persons. Diagram 3 lays out the multitudinous self.

[81] Flanagan, natural method.



Part VI. Conclusion

In this chapter, I filled in some gaps in Hacking's account of looping effects by introducing the multitudinous self in its causal trajectory. In particular, I have argued that there are two connected gaps in Hacking's analysis of looping effects. First, an empirically and philosophically plausible account of the self is missing in the causal structure of looping effects. Second, Hacking fails to engage with the complexity of mental disorder in the consideration of this phenomenon in the realm of psychopathology. Due to these shortcomings, it is not explicit in Hacking's looping effects how exactly classifications of mental disorders change the self-concepts and behavior of those diagnosed with these conditions. I offered an empirically and philosophically plausible model of the self that I call the multitudinous self which fills in these gaps. Multitudinous self, capturing the complexity of selfhood and the encounter

with mental disorder, makes explicit how self-concepts are formed, how they evolve, and how they motivate behavioural changes in the subjects. Grounded as it is in the sciences of the mind and responsive to the experiences of those living with mental disorders, the multitudinous self better explains the causal trajectory of looping effects. Multitudinous self, I further suggested, is a fruitful schema for both the scientific research programs in their investigation of mental disorders and the clinical and ethical contexts in facilitating successful interventions into the lives of those with mental disorders, allowing them to flourish. Thus, with the multitudinous self, I advocate a new style of reasoning about mental disorders in philosophy of psychiatry.

Acknowledgements: I owe special thanks to Jackie Sullivan, Owen Flanagan, Muhammad Ali Khalidi, Harold Kincaid, George Graham, Peter Zachar, David Jopling, and Nathan Brett for their helpful feedback on this chapter. An earlier version was presented at Dalhousie University Philosophy Department colloquium. I am grateful to the audience for their comments. I acknowledge Canadian Institutes of Health Research (CIHR), NNF 80045, States of Mind: Emerging Issues in Neuroethics.

References

Amador, X. F., Seckinger, R. A. 1997. The assessment of insight: A methodological review. *Psychiatric Annals*, 27,798–805.

Amador, X. F., Strauss, D. H., Yale, S. A., & Gorman, J. M.1991. Awareness of illness in schizophrenia. *Schizophrenia Bulletin*, 17, 113–132.

Bartlett, F.C. 1932. *Remembering*. Cambridge University Press.

Bogen, J. 1988. Comments. *Nous*, 22, 65-66.

Bowlby, J. 1969. *Attachment and Loss*. London: Hogarth Press.

Campbell, S. 2003. *Relational remembering: Rethinking the memory wars*. Maryland: Lowman and Littlefield.

Carlson Licia. 2010. *The Faces of intellectual disability: Philosophical Reflections*. Bloomington: Indiana University Press.

Cooper, R. 2007.

Cooper, R. 2004a. Why Hacking is Wrong about Human Kinds? *British Journal of Philosophy of Science,* 55, 73-85.

Cooper, R. 2004b. What is Wrong with the DSM? *History of Psychiatry,* 15 (1), 5-25.

Corrigan, P. W., Edwards, A., Green, A., Diwan, S. E., & Penn,D. L. (2001). Prejudice, social distance, and familiarity with mental illness. Schizophrenia Bulletin, 27, 219–225.

Corrigan, P. W., & Holzman, K. L. 2001. Do stereotype threats influence social cognitive deficits in schizophrenia? In P. W. Corrigan & D. L. Penn (Eds.), Social cognition and schizophrenia (pp. 175–192). Washington, DC: American Psychological Association.

Corrigan, P. W., & Lundin, R. K. 2001. Don't call me nuts! Coping with the stigma of mental illness. Tinley Park, IL: Recovery Press.

Corrigan, P. W., & Nelson, D. 1998. Factors that affect social cue recognition in schizophrenia. Psychiatry Research, 78, 189–196.

Douglas, M. 1986. *How Institutions Think*. Syracuse, NY: Syracuse University Press.

Farina, A. 1998. Stigma. In K. T. Mueser & N. Tarrier (Eds.), Handbook of social functioning in schizophrenia (pp. 247–279). Boston: Allyn & Bacon.

Graham, G. 2010. *The Disordered Mind.* Routledge.

Flanagan, O. 1991. *Varieties of moral personality*. Cambridge, MA: Harvard University Press.

Fogel, A. 1993.*Developing Through Relationships: Origins of Communication, Self, and Culture*. London: Harvester Wheatsheaf.

Gibson, E. 1993. Ontogenesis of the Perceived Self. In *The Perceived Self*, ed. U. Neisser. London: Cambridge University Press.

Gilbert, D. 2006. *Stumbling on happiness*. New York: Vintage Books.

Hacking, I. 2007a. Natural Kinds: Rosy Dawn, Scholastic Twilight. *Royal Institute of Philosophy Supplement*, 61, 203-239.

Hacking I. 2007b. Kinds of People: Moving Targets. *Proceedings of the British Academy,* 151:285-318.

Hacking, I. 2004.Between Michel Foucault and Erving Goffman: Between discourse in the abstract and face-to-face interaction. *Economy and Society,* 33 (3), 277-302.

Hacking, I. 2002. *Mad Travelers: Reflections on the Reality of Transient Mental Illnesses*. Cambridge: Harvard University Press.

Hacking, I. 2000. *The Social Construction of What.* Cambridge, MA: Harvard University Press.

Hacking, I. 1999. Kind Making: The Case of Child Abuse. In *The social construction of what?* Cambridge, MA: Harvard University Press.

Hacking, I. 1995a. *Rewriting the Soul: Multiple Personality and the Science of Memory*. New Jersey: Princeton University Press.

Hacking, I. 1995b. The Looping Effects of Human Kinds. In *Causal Cognition, eds.* D. Sperber and A.J. Premack, 351-383.Oxford: Oxford University Press.

Hacking, I. 1992. World-Making by Kind-Making: Child Abuse for Example. In *How Classification Works, eds.,* M. Douglas and D. Hull, 180-238. Edinburgh: Edinburgh University Press.

Hacking, I. 1986. Making Up People. In *Reconstructing Individualism*, ed. T. Heller, M. Sosna and D. Wellberry, 222-236. Stanford, CA: Stanford University Press.

Jopling, D. A. 2000. *Self-knowledge and the self.* New York, NY: Routledge.

Jopling, D.A. 1997. A "Self of Selves". In *The Conceptual Self in Context*, eds., U. Neisser and D.A. Jopling, 249-267. New York. Cambridge University Press.

Kesebil, Uttal, Gardner 2010. Socialization: Insights from Social Cognition Social and Personality Psychology Compass 4/2 (2010): 93–106

Khalidi, M.A. 2010. Interactive Kinds. *British Journal of Philosophy of Science*, 61, 335 – 360.

Kosslyn, S. 2006. On the evolution of human motivation: The Role of Social Prosthetic Systems. In *Evolutionary Cognitive Science*. Shackelford, T., and Keenan, J. eds. Cambridge: MA. MIT Press, 541-554.

Marin, K., Bohanek, J. G., & Fivush, R. 2008. Positive effects of talking about the negative: Family narratives of negative experiences and preadolescents' perceived competence. *Journal of Research on Adolescence, 18* (3) 573-593.
Murray, L. and Trevarthen C. 1985. The infant in mother-infant communication. *Journal of Child Language*, 13, 15-29.

Miller, P. J., Potts, R., Fung, H., Hoogstra, L., & Mintz, J. 1990. Narrative practices and the social construction of self in childhood. *American Ethnologist, 17*, 292-311.
Neisser, U. 1988. Five kinds of self-knowledge. *Philosophical Psychology*, 1, 35- 59.

Neisser, U. and Jopling, D. 1997. The conceptual self in context: Culture, Experience, Self-understanding. Cambridge: Cambridge University Press.

Neisser, U and Fivush, R. 1994. The Remembering Self: Construction and Accuracy in the Self-nattarive. Cambridge: Cambridge University Press

Neisser, U.1993. The perceived Self. Ecological and Interpersonal Sources of Self- Knowledge. New York: Cambridge University Press

Nisbett, R. & Wilson, T. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Pennebaker, J. W. 1993. Putting stress into words: Health, linguistic, and therapeutic implications. *Behavior Research & Therapy, 31*(6), 539-548.

Pickering, J. 1999. The Self is a Semiotic Process. In *Models of the Self*, eds. Gallagher, S. and Shear, J., 63-79.

Phelan, J., Link, B., Stueve, A., & Pescosolido, B. (2000). Public conceptions of mental illness in 1950 and 1996: What is mental illness and is it to be feared? Journal of Health and Social Behavior, 41, 188–207.

Saks, E. 2007. *The center cannot hold: My journey through madness*. New York: Hyperion.

Samuels, R. 2009. Delusion as a Natural Kind. In *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, eds., M. Broome and L. Bortolotti, 49-82. L. Oxford, UK: Oxford University Press.

Stern, D. 1993.The role of feeling for an interpersonal self. In *The Perceived Self*. Ed. U.Neisser. London: Cambridge University Press.

Stets, J.E.; P.J. Burke. 2003. A Sociological Approach to Self and Identity. In *Handbook of Self and Identity*, eds., M. Leary and J. Tangney, 128-152. NY: The Guilford Press.

Tekin, Ş. Forthcoming. Self-insight in the time of mood disorders: After the diagnosis, beyond treatment. *Philosophy, Psychiatry, Psychology*.

Tekin, S. 2011. Self-concepts through the diagnostic looking glass: Narratives and Mental Disorder. *Philosophical Psychology*, 24:3, 357 – 380.

Tekin, S. 2010. Mad Narratives: Self-Constitutions Through the Diagnostic Looking Glass. PhD Dissertation, York University, Toronto, Canada.

Trevarthen, C. 1980. The foundations of intersubjectivity: Development of interpersonal and cooperative understanding in infants. In *The social foundations of language and thought,* ed., D. Olson, 316-342. New York: Norton.

Wahl, O. F. (1999). Mental health consumers' experience of stigma. Schizophrenia Bulletin, 25, 467–478.

Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, Massachusetts: Harvard University Press.

Zachar, P. 2001. Psychiatric disorders are not natural kinds. *Philosophy, Psychiatry, and Psychology*, 7 (3), 167-182.

The Relaxed Forces Strategy for Testing Natural State Theories: The Case of the *ZFEL*[1]

Derek Turner
Connecticut College
New London, CT, USA
Derek.turner@conncoll.edu

---

1. Introduction. In their recent book, *Biology's First Law*, Dan McShea and Robert Brandon say that they are trying to bring about a "fundamental gestalt shift in how we view evolutionary theory" (2010, p. xi, Cf. p. 128). Their proposed Gestalt shift involves a resetting of the default, or zero-force expectation concerning structural complexity. By "structural complexity" they mean something like internal variance, or internal heterogeneity of parts. They are not talking about functional or adaptive complexity. According to what we might call the received view, stasis (i.e. no change in structural complexity) is the natural state of an evolutionary system, and any directional changes, such as complexity increase, need to be explained by invoking evolutionary forces such as selection. According to McShea and Brandon's zero-force evolutionary law (or ZFEL), the default expectation is that complexity will increase over time. So if we see certain lineages—such as, say, bacteria—that persist for a very long time with no complexity increase, we might explain that on the hypothesis that there is selection against greater complexity. This, the proposed Gestalt shift turns the ordinary understanding of the relationship between selection and complexity on its head.

In this paper, I examine one initially promising strategy, which I call the relaxed forces strategy, for using empirical evidence to discriminate between rival natural state theories. The first step is to identify a case in which one has independent evidence that the relevant evolutionary forces are inoperative. Then one checks to see if the default expectation specified by the natural state theory in question is indeed satisfied under the relaxed forces condition. In section 2, I show how Darwin used this strategy to defend his

1This paper derives from comments on *Biology's First Law* that I gave at an author-meets-critics session at the 2011 ISHPSSB meeting in Salt Lake City, Utah. I am especially grateful to Robert Brandon, Chris Haufe, and Dan McShea for their thoughtful responses to the comments, and for helpful conversation about the *ZFEL*. Thanks also for helpful comments from an audience at POBAM 2012, in Madison, Wisconsin.

Malthusian claim that the natural state of any biological population is geometrical growth. In section 3, I go on to show how McShea and Brandon use this strategy to help build their empirical case in favor of the ZFEL.

I argue, however, that the relaxed forces strategy suffers from a serious defect. The problem is that what counts as a relaxed forces condition depends on the natural state theory that one already holds. It is not possible to identify a relaxed forces condition without presupposing the very theory that one wants to put to the test. I contend, in section 4, that this circularity problem means that the relaxed forces strategy cannot discriminate between rival natural state theories. Darwin's main argument for his Malthusian theory of population is circular. This problem also weakens McShea and Brandon's empirical case for the ZFEL. In section 5, I consider some possible responses to the failure of the relaxed forces strategy.

2. Darwin's use of the relaxed forces strategy. Consider an example of two rival natural state or inertial state theories about population size:

The first theory says that the natural state of any biological population is to remain stable at a given size. If the population has *n* members at a time, and if no external forces or pressures impinge upon it, then it will still have *n* members at the end of some specified time interval. If the population grows or shrinks, then the change in the population's size is to be explained in terms of external forces acting upon it. The default expectation is no change in population size. Deviations from that expectation need to be explained.[2]

The second theory—Darwin's—says that the natural state of any biological population is geometrical growth. If the population has *n* members at a time, and if no external forces or pressures act upon it, then it will have 2*n* members at the end of a specified time interval. Thus, a population that begins with 2 individuals will contain 4 at the end of the first time interval, eight at the end of the next interval, and so on. If the population grows at a slower rate or not at all, that deviation from what is expected is to be explained in terms of external ecological forces.

These two theories have the same structure, but they introduce different default expectations. They make different claims about which is the inertial state of the population. We cannot easily discriminate between the two theories by studying actual populations in nature. Suppose we observe a population that's growing, but at a rate much slower than the geometrical rate that is, in some sense, "predicted" by the second theory. Both theories need to explain what's going on, but the explanations will run in opposite directions. The first theory needs to explain why the population is growing at all; the second needs to explain why it is growing more slowly than expected. In principle, both theories can handle any observed rate of growth, but they will do so in different ways, by appeal to different packages of external forces.

2 Compare Sober (1980, pp. 360 ff.) on the general structure of natural state models.

Darwin knew well that actual populations in nature rarely, if ever grow geometrically, and that if they do, it's only for a short while under special conditions. Virtually every population that we can observe violates the default expectation of his Malthusian theory of population growth. But far from treating observed populations as evidence against his theory, Darwin saw them as an opportunity to put his theoretical machinery to work. He offers a rich catalogue of external ecological forces—predation, disease, food scarcity, and so on—that serve as checks to population increase.

There is an attenuated sense of "predict" in which a natural state theory predicts whatever it specifies as the default expectation. Thus, we might say that Darwin's Malthusian theory predicts that populations will grow geometrically. This usage creates the impression that we can test the theory in question by checking to see whether the default expectation—the "prediction"—is satisfied. This usage is somewhat misleading, however. Nonconformity with the default expectation does not in itself constitute evidence against a natural state theory. It is, rather an opportunity to deploy the explanatory machinery of that theory.

What then (if anything) would constitute evidence against a natural state theory? The *relaxed forces strategy* provides one initially promising answer to this question. Suppose we have independent evidence that the various external forces that may act upon a system are relaxed, so that the system is in something approximating the zero-force condition. In such cases, where we already know that the external forces are relaxed, we can then look to see whether the system conforms to the default expectation. If not, then that would be evidence against the natural state theory. In other words, violations of the default expectation do constitute evidence against the theory, but only in those cases where we already know that the relevant external forces are relaxed. [add: Natural state theories do seem to generate testable predictions about what the system will do when external forces are relaxed.]

In the *Origin of Species*, Darwin at one point seems to use the relaxed forces strategy to argue for his Malthusian principle that "each organic being is striving to increase at a geometrical ratio" (1859, pp. 78-79).

> But we have better evidence on this subject than mere theoretical calculations, namely, the numerous recorded cases of the astonishingly rapid increase of various animals in a state of nature, when circumstances have been favorable to them during two or three following seasons. Still more striking is the evidence from our domestic animals of many kinds which have run wild in several parts of the world: if the statements of the rate of increase of slow-breeding cattle and horses in South America, and latterly in Australia, had not been well authenticated, they would have been quite incredible … (1859, pp. 64-65).

Darwin's reference to "favorable circumstances" suggests that he has in mind cases where the ordinary ecological checks to population growth are relaxed. The cases of introduced species that he proceeds to discuss are also plausibly cases in which many checks to increase, such as food scarcity and predation, are relaxed. Darwin's argument seems to be that in these cases where the forces limiting population growth are relaxed, rare though such cases may be, we do in fact typically see growth at the expected geometrical rate.

Few biologists today accept Darwin's Malthusian claim that geometrical growth is the natural state of any biological population. This Malthusian view is no longer regarded as essential to the theory of natural selection. Darwin himself, however, thought that he had some empirical support for his preferred natural state theory. Insofar as biologists no longer find Darwin's argument too convincing, it's worth asking why. Is there some problem specific to Darwin's own use of the relaxed forces strategy? Or is there some deeper problem with the strategy itself? In section 4, I will argue that the latter is indeed the case.

Newtonian mechanics is another familiar example of a natural state theory. (Interestingly, McShea and Brandon draw an explicit analogy between the ZFEL and Newton's theory [2010, p. 6].) The relaxed forces strategy clearly would not work well as a test of Newton's theory. We'd have to identify an object that is not being acted upon by any physical forces, and then check to see whether the object exhibits uniform rectilinear motion. The problem here seems like a practical one: the relaxed forces condition seems impossible to bring about, or even to approximate, because the interfering forces are always operative. In other contexts, however, such as those involving evolutionary systems, we might really be able to observe a system in the relaxed forces condition.

3. McShea and Brandon's predictive test of the ZFEL. Much as Darwin argued that biological populations have a natural tendency to grow at a geometrical rate, so McShea and Brandon argue that all evolutionary systems naturally tend toward greater structural complexity/heterogeneity. They readily acknowledge that the ZFEL is not an ordinary empirical generalization of the form, *All F's are G's* (2010, p. 7). They do insist, however, that it is empirically testable:

> The ZFEL is not analytic. It is not true as a matter of logic or mathematics, as is biology's so-called Hardy-Weinberg Law. Rather, it is synthetic, making an empirical claim about the way the world is (2010, p. 6).[3]

3 McShea and Brandon's views about the status of the ZFEL are quite nuanced. They say that the ZFEL is a synthetic, empirical claim, but then they also say, elsewhere, that it is "ultimately reducible to probability theory," which at first blush would seem to suggest that the ZFEL is analytic. But McShea and Brandon also hold that "there is a bit of probability theory that is not pure math" (2010, p. 109).

In the course of the book, they argue that there is quite a lot of empirical evidence for the ZFEL, and they propose some empirical tests of it. Like Darwin, they use the relaxed forces strategy in order to build their empirical case for the ZFEL.

McShea and Brandon take care to define the ZFEL in such a way that it only applies to biological systems:

> ZFEL (special formulation): In any evolutionary system in which there is variation and heredity, in the absence of natural selection, other forces, and constraints acting on diversity or complexity, diversity and complexity will increase on average (2010, p. 3).

Strictly speaking, the ZFEL does not apply to systems that do not exhibit variation and heredity. McShea and Brandon do at least consider a more generalized version of the ZFEL—the G-ZFEL—that would apply to all physical systems (p. 112).[4] They use the example of a freshly painted picket fence to illustrate the ZFEL. Strictly speaking, though, the picket fence would only be an instance of the G-ZFEL, and not of the ZFEL proper, because the picket fence is not "an evolutionary system in which there is variation and heredity."

To begin with, all the pickets in the freshly painted fence look more or less the same; there is very little internal differentiation among the parts of the fence. Over time, however, the different pickets come to exhibit different patterns of weathering. The paint peels a little more on those most exposed to the sun. One or two of the pickets begin to rot away where they make contact with the ground. One is knocked loose by a soccer ball. Over time, what began as a homogeneous system becomes more and more heterogeneous. The fence gets more complex, in McShea a Brandon's technical sense. It's tempting to say—and McShea and Brandon do say—that in this example, increasing heterogeneity is what we should expect. If the fence remained homogeneous for a long time, that would be surprising and would need explanation. Maybe someone repaints it periodically. That, at least, is the kind of external force we'd have to invoke in order to explain why heterogeneity does not increase.

Even here, though, with this lovely stage-setting example, it's not clear that any evidence compels us to set the default expectation one way rather than another. One could also formulate a rival natural state theory in which stasis—i.e., no change in internal variance—is the default expectation and deviations from that are what need to be explained. Complexity increase, for example, would need to be explained in terms of the differential action of external forces—sun, wind, rain, soccer balls, and the like—on the pickets. The general point is that we seem to have some flexibility in determining what the default expectation is going to be, and hence some flexibility in deciding which

4 This generalized version of the ZFEL is a descendant of Herbert Spencer's law of the instability of the homogeneous.

natural state theory to work with. It's not clear that any actual observations of the fence itself will tell us what the natural state is—whether it is increasing complexity or stasis. If we watch the fence, we'll see that complexity does in fact increase, but both natural state theories can handle that observation equally well.

In a recent review of McShea and Brandon's book, Matthen (2011) describes what he seems to think is a counterexample to the generalized ZFEL (or G-ZFEL). Imagine a large mural on a wall, done in the pointillist style of Seurat. The mural has a huge amount of internal variance; each dot is slightly different from the others. Over time, the colors fade and the heterogeneity of the system diminishes. This increasing homogeneity seems just the opposite of what the ZFEL says we should expect.

Matthen, in fact, does more than merely suggest that the fading pointillist mural is a counterinstance of the ZFEL. He wants to make a deeper point that whether complexity is increasing or decreasing may well depend on our level of description of the mural. It's most natural, perhaps, to say that the fading of the mural involves decreasing internal variance. But what if we focused instead on the spatial locations of the molecules of blue pigment. Let's imagine that there is just one bright blue region in the freshly painted mural. At the beginning of the process, the molecules of blue pigment have a low degree of spatial variance; they're all in the same area. As the painting fades, and the molecules of pigment weather off, their spatial variance increases. Many of them end up out in the environment. And that's an increase in complexity in McShea and Brandon's sense.

There is a more straightforward point here that Matthen might have made but doesn't. Even setting aside Matthen's argument concerning levels of description, it isn't entirely clear why anyone would think that the fading pointillist mural would be a counterexample to the ZFEL in the first place. Yes, the fading mural is a system that violates the default expectation of increasing internal variance. This just means that we have to invoke external forces—such as the sun and the elements—in order to explain the surprising decrease in complexity this case. A deviation from the default expectation of a natural state theory is not a counterexample to that theory. It is, rather, an occasion to put the theory to work by invoking external forces.

Part of the appeal of the relaxed forces strategy is that it seems to permit novel predictive tests of natural state theories. Although there is quite a bit of debate about what makes a prediction novel, and why novelty should confer any extra evidence over and above the evidence we already get from showing that a theory has true observational consequences, some philosophers think of novel predictive success as the very best sort of evidence that one can give for a theory (see, e.g. Leplin 1997). And it seems possible to make novel predictions about what one will observe in cases where the external forces are relaxed.

McShea and Brandon argue that the ZFEL passes at least one interesting novel predictive test, and they propose some other tests in the same vein. I want to focus here on the test that they showcase a bit (2010, pp. 73-76).

Consider the following experiment: Subject male mice to doses of radiation, and then look at the effects on the vertebral columns of their offspring. "The ZFEL prediction is that, absent selection and constraint, offspring will tend to be morphologically more complex than their parents" (2010, p. 73). One appealing thing about focusing on the vertebral column is that it is relatively straightforward to arrive at a measure of complexity. The mammalian vertebral column lends itself to division into part types (cervical, thoracic, lumbar, etc.). In the imagined experiment, one can come up with a list of possible changes in the rat vertebrae that would all count as increases in complexity in McShea and Brandon's technical sense. These changes include:

- Dyssympysis, in which to pieces of a vertebra fail to fuse property during development.
- The total absence of one part of a vertebra, such as a neural arch.
- The duplication of one part of a vertebra.
- The fusion of two adjacent vertebra.
- Any malformation or change in size of one vertebra.

After irradiating the male mice, study the complexity of the vertebral columns of their offspring. How many of the changes observed led to increases in complexity, and how many led to decreases in complexity? Examining the offspring mice as newborns is a way of making sure that selection has no chance to operate (although McShea and Brandon acknowledge that there is some differential survival before birth), thus making this a relaxed forces case. Does complexity increase when selection is relaxed?

It turns out that it does. In the 1960s, some scientists at Oak Ridge National Laboratory actually carried out the above procedure, though obviously not with the aim of testing the ZFEL (Ehling 1965). They divided the observed morphological changes into two groups. The class I changes occurred in just one animal, while the class II changes occurred in more than one animal. The researchers observed 20 class I changes in 10 animals, and 16 class II changes. After going back and reviewing Ehling's study, McShea and Brandon determined that no fewer than 17 of the class I changes involved increases in structural complexity, while all of the class II changes did. The ZFEL predicts that when the relevant external forces (especially selection) are relaxed, complexity will increase. That prediction was clearly born out in this case. What's more, the prediction is a genuinely novel one in the epistemic sense, which is the sense that counts. The ZFEL was not in any way tailored to accommodate these results. Even though the experiment was done back in the 1960s, McShea and Brandon presumably didn't formulate the ZFEL with these results in mind.

4. A circularity problem for the relaxed forces strategy. The relaxed forces strategy requires that we be able to specify in advance which cases are the ones where the forces are relaxed. We are then supposed to look at those cases to see whether the default, or zero-force expectation is at least approximated. But how do we know what counts as a

relevant interfering force in the first place? Rival natural state theories make different claims about which external forces are the relevant ones.

The picket fence affords a nice illustration of the circularity problem that afflicts the relaxed forces strategy. Suppose we want to test the theory (the G-ZFEL) which says that complexity increase is the natural state for a picket fence. In order to do that, we try to identify a case where all the external forces that might act upon the fence are absent or at least relaxed. What forces are those? Should we, for instance, try to find a fence that is not impacted by any stray soccer balls? No. Since soccer ball impacts serve only to increase the complexity of the fence—say, by knocking loose one picket but not the others—the stray soccer ball would not count as an external force in this case. What makes a force external is precisely the fact that it works to keep the system out of its natural or inertial state. Here a painter who re-fastens loose pickets and repaints the fence periodically would count as an external force. In order to know which forces are external, we have to know which is the natural state of the system. In order to identify cases where the external forces are relaxed, we have to know which forces are the external ones. For this reason, the relaxed forces strategy requires us to make assumptions about the natural state of the system. It requires us to assume the very claim that we thought we were testing.

To make the above point more vivid, imagine a proponent of a rival natural state theory—say, the one according to which stasis (or no change in complexity) is the natural state of the picket fence. The advocate of the stasis theory would classify stray soccer balls as external forces, because they obviously work to keep the fence out of its natural state. So the proponent of the stasis theory and the proponent of the G-ZFEL would just disagree about which cases are the ones in which the external forces are relaxed.

This circularity problem also undermines McShea and Brandon's novel predictive test of the ZFEL. Again, one way to bring this out is to try to envision the response that an advocate of a different natural state theory might make. Consider how this experiment might look to a defender of the stasis theory. According to that view, the default expectation is that no change will occur in the complexity of the mouse vertebral columns from one generation to the next. In this case, complexity obviously increased. So on the stasis view, that deviation from the default expectation needs to be explained in terms of the operation of some external forces. Which external forces might those be? Selection will not do the trick here, because as McShea and Brandon point out, the experiment is set up in a way that guarantees reduced selection. The stasis theorist, however, has a different answer at the ready: the external forces that explain the complexity increase are just the scientists who irradiated the male parents. But for the radiation, the vertebral columns of the offspring mice would probably have exhibited about the same degree of complexity as those of their parents. Indeed, the stasis theorist could propose a relaxed forces test of her own: Just measure the complexity of the offspring's vertebral columns without dosing the male parents with radiation.

McShea and Brandon do not treat the scientists who dose the male mice with radiation as an external force operating on the system. That's just because, given the natural state theory they defend—namely, the ZFEL—the external forces are, by definition, the ones that work against complexity increase. Since the scientists are causing complexity increase, they don't count as an external force. This parallels the example of the picket fence and the soccer ball quite nicely. Stray soccer balls don't count as external forces if we are assuming that complexity increase is the zero-force condition for the picket fence, because stray soccer balls do not work against complexity increase.

In order to identify a relaxed forces case—or more precisely, in order to show that the case of the irradiated mice *is* one in which the external forces are relaxed—McShea and Brandon must presuppose the ZFEL. They must presuppose the natural state theory that they purport to be putting to the test. Someone who prefers a different natural state theory may simply deny that this particular case is a relaxed forces case. Nor is this problem an isolated one. It affects some of the other tests they propose. For example, they suggest that it might be illuminating to study the complexity of structures that are not under selection, such as the eyes of cave-dwelling crayfish (2010, pp. 76-77). Do cave-dwelling crayfish have more complex eyes than their surface-dwelling relatives? If so, then that would mean that complexity increases in the relaxed forces condition. Here again, it might be open to a proponent of a rival natural state theory to challenge the claim that the relevant external forces are really relaxed in the case of the cave-dwellers.

Darwin's argument for the Malthusian natural state theory of population growth (section 2) also runs afoul of this circularity problem. Darwin's idea was to test the claim that populations naturally tend to grow at a geometrical rate of increase by looking at situations in which the usual checks to population growth are relaxed—for example, cases in which a newly introduces species has abundant food and no predators. However, someone who accepts a different natural state theory at the outset—say, someone who thinks that the natural state of populations is stasis with respect to size—would presumably not identify the interfering forces in quite the same way. On this rival theory, predation might count as an interfering force (for example, if it caused a reduction in the population size, and hence a deviation from the expected stasis), but it also might not. On this rival theory, the relaxed forces condition could be one in which there is significant predation.

This circularity problem means that the relaxed forces strategy cannot discriminate empirically between rival natural state theories. The problem, in a nutshell, is that the strategy assumes that we already know what to count as the relevant interfering forces. But that is precisely what we are trying to find out. There might, however, be other ways of subjecting natural state theories to empirical assessment. And even if there were not—that is, even if the choice between rival natural state theories were underdetermined—such theories might still have an indispensable role to play in scientific inquiry. The failure of the relaxed forces strategy raises some larger questions

about the status of natural state theories, and I will not be able to address those questions fully here. In the next section, however, I'll try to advance the discussion by considering some possible responses that McShea and Brandon could make to the failure of the relaxed forces strategy.

5. Possible responses to the problem. How might McShea and Brandon continue to defend the ZFEL in light of the alleged failure of the relaxed forces strategy? The most natural response to the failure of the relaxed forces strategy would be to seek out some *other* means of subjecting rival natural state theories to empirical tests. While I would not want to rule out the possibility that some other strategy might be more promising, I confess that it is difficult to imagine what such a strategy might look like. In order to test a natural state theory, we must examine either a relaxed forces case or a non-relaxed forces case. (Every case falls into one of those two categories.) If it doesn't help to look at a relaxed forces case, why would it help to look at a non-relaxed forces case?

In fairness to McShea and Brandon, the putative novel predictive test that I described in section 3 is just one small piece of the larger argument they make in favor of the ZFEL. It's not clear that my critique of that "test" should derail the larger project. McShea and Brandon repeatedly stress the unifying or explanatory power of the ZFEL:

> So what is the point of the ZFEL? First, it offers unity. A heretofore unconnected set of phenomena … are revealed to be instances of the same underlying principle (2010, p. 71).

Indeed, in the course of the book, they show how a wide variety of biological phenomena —from increasing disparity in macroevolution (pp. 36-38), to the observation that characters not under selection show greater variation (p. 70) – all seem to fall into place when viewed through the lens of the ZFEL. The crucial question, though, is perhaps not whether the ZFEL does a good job unifying biological phenomena that were "heretofore unconnected," but whether it does a better job unifying those phenomena than a rival natural state theory, such as the stasis theory for diversity/complexity, would do. Darwin might provide a helpful comparison here: It is fairly easy to see how Darwin's picture of populations "striving" to increase in the face of myriad constraints and checks can unify a wide variety of biological phenomena. The real question is whether it does a better job unifying those phenomena than the rival stasis theory would do. Assessing this appeal to the unifying power of the ZFEL is a larger project than I can take on here.

A second possible response to the failure of the relaxed forces strategy is to make a subjectivist move. McShea and Brandon briefly flirt with such a move, but it is difficult to make out how much sympathy they really have for it (pp. 102-103). When we talk about the natural state of a system, we seem to be talking about the system itself. However, when we talk about expectations, we seem to be talking about ourselves. Expectation and surprise are merely subjective, psychological notions. One possibility is that there is nothing at issue here above and beyond what we expect, and hence no deep mind- or theory-independent fact of the matter about what the natural state of a system

"really" is. We might expect biological populations to grow geometrically, or we might expect them to remain stable. When we talk about populations as having a natural state, we are merely reading our own expectations into the natural world. Which observations will surprise us and will seem to require explaining will also depend on our initial expectations. But there may be no fact of the matter about which expectations are the right ones to have. Moreover, each natural state theory picks out a set of relevant interfering forces or constraints that can make a difference to a system. There may be no fact of the matter about which is the right way to identify those interfering forces.

At one point in the book, McShea and Brandon appear to make this subjectivist move:

> [A]re there objective matters of fact that settle what count as forces in a particular science, and so what counts as the zero-force condition, or is this a matter of how we set out our theory, and so a matter of convention? (2010, p. 102)
>
> We will not dare to try to answer this question in general, though we will share our suspicions: in some cases objective facts will settle the matter, but in most cases they will not. But in the present case it is clear that we must take a conventionalist stance … (2010, p. 103).

McShea and Brandon acknowledge that their decision to treat complexity increase as the natural state of any evolutionary system is a "choice," and in the passage immediately following the one quoted above, they give some reasons for this choice. They point out that they have decided to focus narrowly on systems that involve reproduction, variation, and heritability. (Indeed, the special formulation of the ZFEL restricts its application to systems that exhibit variation and heritability). What's not entirely clear is why the decision to focus on systems that exhibit reproduction, variation, and heritability would give us a reason to choose to adopt complexity increase (rather than, say, stasis) as our default expectation for evolutionary systems.

The subjectivist move may also be somewhat at odds with the larger project of McShea and Brandon's book. One of their goals, as we have seen, is to build an empirical case for the ZFEL. But if one became convinced that there is no deep fact of the matter about whether the natural state of evolutionary systems is complexity increase vs. stasis—if one thought that the difference between these rival pictures is merely a matter of our subjective expectations—then the project of building an empirical case for one natural state theory over the other would seem unmotivated. If subjectivism is correct, then the most we can do is try to persuade others to change their expectations. Thus, although they do seem to sympathize somewhat with this subjectivist move, it may be more charitable to read McShea and Brandon as holding that there really is an objective fact of the matter about whether the ZFEL is correct.

To sum up the results of this section: There are at least three possible ways in which McShea and Brandon could respond to the failure of the relaxed forces strategy. (1) They might seek some other means of testing the ZFEL empirically, though it's difficult at this point to see what such a strategy might look like. (2) They might appeal to the unifying power of the ZFEL, though it would also be necessary to show that the ZFEL has greater unifying power than rival natural state theories. Finally (3) they could make a subjectivist move and abandon the assumption that there is a mind- and theory-independent fact of the matter about which natural state theory is true of a given system. This all too brief discussion is just the beginning of an attempt to map out some of the logical space of possible responses. Most importantly, the failure of the relaxed forces strategy should be an occasion to revisit questions about the status of natural state theories and their role in empirical science.

6. Conclusion. Natural state theories have figured prominently in the history of philosophy, from Aristotle to Spinoza. They also show up repeatedly in natural science, from Newtonian mechanics to the idea that the Hardy-Weinberg equilibrium describes the natural state of biological populations. In this paper, I have examined one initially promising strategy for subjecting such theories to empirical tests. The relaxed forces strategy was employed by Darwin in the *Origin*, and it has been revived more recently by McShea and Brandon. I've argued that the strategy fails to deliver the goods, because it is plagued by a circularity problem. The failure of this strategy weakens McShea and Brandon's empirical case for the ZFEL, but (as I argued in section 5) need not derail their project entirely. Nevertheless, the failure of the relaxed forces strategy should prompt philosophers to think further about the status and role of natural state theories in science.

As far as I know, few biologists today think that we even need a natural state theory for population size. It's not that biologists have abandoned Darwin's Malthusian picture in favor of some rival natural state theory—say, a stasis theory. Rather, they have abandoned it in favor of no natural state theory at all, and they have done so in spite of the unifying power of the Malthusian view. Trends in population size are phenomena to be explained, but modern evolutionary theory makes no assumptions about natural tendencies in population size. McShea and Brandon do show how powerful a natural state theory of diversity/complexity can be. There is also a deeper issue here about which biological phenomena call for natural state theorizing and which do not. The failure of attempts to test natural state theories just brings that question into sharper relief.

References

Darwin, C. (1859/1964) *On the Origin of Species (A Facsimile of the First Edition)*. Cambridge, MA: Harvard University Press.

Leplin, J. (1997) *A Novel Defense of Scientific Realism*. Oxford: Oxford University Press.

Matthen, M. (2011) Review of Biology's First Law: The Tendency for Diversity & Complexity to Increase in Evolutionary Systems. *Notre Dame Philosophical Reviews*. Available online at http://ndpr.nd.edu/news/24573-biology-s-first-law-the-tendency-for-diversity-and-complexity-to-increase-in-evolutionary-systems/. Last accessed 18 Mar 2012.

McShea, D., and R. Brandon (2010) *Biology's First Law*: *The Tendency for Diversity & Complexity to Increase in Evolutionary Systems*. Chicago, IL: University of Chicago Press.

Sober, E. (1980), "Evolution, population thinking, and essentialism," *Philosophy of Science* 47(3): 350-83.

THE PROBLEM OF PHILOSOPHICAL ASSUMPTIONS AND CONSEQUENCES OF SCIENCE

Jan Woleński

Abstract: This paper argues that science is not dependent on philosophical assumption and does not entail philosophical consequences. The concept of dependence (on assumptions) and entailment is understood logically, that is, are defined via consequence operation. Speaking more colloquially, the derivation of scientific theorems does not use philosophical statements as premises and one cannot derive philosophical theses from scientific assertions. This does not mean that science and philosophy are completely separated. In particular, sciences leads to some philosophical insights, but it must be preceded by a hermeneutical interpretation.

It is frequently asserted that science assumes some philosophical premises or/and leads to philosophical consequences. For instance, transcendental epistemologists (Kant, Neo-Kantians) argue that epistemology establishes conditions of validity for any kind of cognition, including scientific one. According to Kant, every experience locates its objects in space and time. Thus, assertions about space and time, more specifically that space is three-dimensional and time is absolute, belong to philosophical presuppositions of science. Husserl expressed a similar view, although oriented more ontologically than epistemologically, particularly strongly (italic in the original):

> If, however, all eidetic science is intrinsically independent of all science of fact, the opposite obtains, on the other hand, in respect of *the science of fact* itself. *No fully developed science of fact could subsist unmixed* with eidetic knowledge, and in consequent *independence of eidetic science formal or material.* For *in the first place* it is obvious that an empirical science, wherever it finds grounds for its judgments through mediate reasoning, must proceed according to the *formal* principles used by formal logic. And generally, since like every science it is directed towards objects, it must be bound by the laws which pertain to the essence of *objectivity in general. Thereby it enters into relation with the group of formal-ontological* disciplines, which, besides formal logic in the narrower sense of the term, includes the disciplines figured formerly under the formal "*mathesis universalis*" (hus, arithmetic also pure analysis, theory of manifolds). Moreover, and in *the second place*, every fact includes

an essential factor of a *material* order, and every eidetic truth pertaining to the pure essence thus included must furnish a law that binds the given concrete instance and generally every possible one as well.

[…]

Every factual science (empirical science) *has essential theoretical bases in eidetic ontologies.* […] In this way, for instance, the eidetic science of physical nature in general (the *Ontology of nature*) corresponds to all the natural science disciplines, so far indeed as an Eidos that can apprehended in its purity, the "essence" *nature in general*, with an infinite wealth of included essential contents, corresponds to actual nature.[1]

Husserl ascribes to formal ontology a very essential role, because, according to him, all factual (empirical) assertions have their ultimate basis in fundamentals established by eidetic analysis.

Another frequently explored link between science and philosophy consists in looking for philosophical consequences of scientific theories or even singular scientific theorems.[2] Mathematics provides a very good example in this respect. Some people maintain that classical mathematics implies Platonism, although others regard antirealism as a consequence of constructive mathematics. Passing to physics, Newtonian mechanics is reputed to entail determinism, but indeterminism is qualified as having its inferential foundation in quantum theory; this connection will be exploited several times in this paper. Similarly, vitalism is considered as following from embryology as a part of biology, although theory of evolution goes together with mechanism as its philosophical output. Gödel's incompleteness theorems are sometimes taken as premises in arguments for non-reducibility of mind to machines. Another use the same metamathematical results consists in attempts to show that knowledge is essentially uncertain. There is a good example:

[…] I single out for discussion – the question whether the law of excluded middle, when it refers to statements in the future tense, forces us into a sort of logical Predestination. A typical argument is this. If it is true now that I shall to do a certain thing tomorrow, say to jump into the Thames, then no matter as fiercely I resist […], when a day comes I cannot help jumping into the water; whereas, if this prediction is false now […] it is impossible for me to spring. Yet that the prediction is either true or false is itself a necessary truth, asserted by the law of excluded middle. From this the startling consequence seems to follow […] that indeed the entire future is somehow fixed, logically preordained.[3]

[1] E. Husserl, *Ideas. General Introduction to Pure Phenomenology*, tr. by W. F. Boyce Gibson, Collier Macmillan, London 1931, p. 57/58. For more recent similar statements see, for example, I. Stein, *The Concept of Object as the Foundation of Physics*, Peter Lang, Frankfurt am Main 1996 or M. Heller, *Ultimate Explanations of the Universe*, Springer, Berlin 2009. See also notes 2an d 5.

[2] See F. Weinert, *The Scientist as Philosopher. Philosophical Consequences of Great Scientific Discoveries*, Springer, Berlin 2004 for a historical survey. I choose this book for its subtitle clearly related to the problem announces by the title of the present paper. General and special literature about philosophical consequences (and assumptions as well) of science is enormous. In fact, every textbook of philosophy of science or monograph in this area addresses to this topic directly or indirectly. See note 5 for an additional selected bibliography.

[3] F. Waismann, "How I See Philosophy", in F. Waismann, *How I See Philosophy*, Macmillan, London 1968, p. 8/9. Note that Waismann himself does not accept the argument from the excluded middle to logical Predestina-

Social sciences and humanities also share philosophical import with natural disciplines (*the* science in the traditional science), although one should notice that strict borderlines between philosophical and non(or less)-philosophical regions are difficult to depict them univocally. We easily observe that the relation between science and philosophy is less and less explicit if we go to further members in the sequence {mathematics, physics, chemistry, biology, social sciences, humanities}. By the way, this succession is almost identical with Comte's classification of abstract sciences. In order to simplify my considerations, I will entirely omit philosophical problems of social sciences and humanities, and limit discussion about formal sciences (logic and mathematics) to some illustrative examples. Thus, I focus on natural science, mostly physics.

I will try to introduces some conceptual order into the problem of philosophical assumption and consequences of science. The issue in question require some clarifications for several reasons. In general and to anticipate my position, I will argue that science does not need philosophical assumption as well as it does not have philosophical consequences. Yet this view does not imply that science and philosophy are mutually independent. On the contrary, science suggests a lot of philosophical problems and perhaps could lead to philosophical solutions, although the later hope should be taken modestly and with various additional constraints (I will return to this question at the end of this paper). The reverse dependence, that is, an influence of philosophy on science, is a much more delicate matter, although explicit philosophical roots of several scientific discoveries (for example, Platonic background of Copernicus' theory) are very well confirmed by the history of science. In fact, historical studies seem to suggest that the role of philosophy as a source of scientific results continuously weakens through the course of time. Anyway, we need to distinguish the question whether there are philosophical problems of science from the issue whether science has philosophical assumptions and leads to philosophical consequences. The lack of this distinction obscures any analysis of the problem in question. And this is the first motive for trying to do a clarifying work.

Secondly, philosophers and scientists are not always clear whether they speak about philosophical assumptions of science or its philosophical consequences. Let me illustrate this once again by the relation of logic to determinism and indeterminism:

The law of bivalence is bivalence is the basis of our entire logic, yet it was already much disputed by the ancients. Known to Aristotle, although contested for propositions referring to future contingencies; peremptorily

tion. Thus, Waismann's text should be regarded as a reconstruction of an argument proposed by someone else (see below).

rejected by Epicureans, the law of bivalence makes its full appearance with Chrysippus and the Stoics as a principle of their dialectics, which represents the ancient propositional calculus […]. The quarrel about the law of bivalence has a metaphysical background, the advocates of the law being decides determinists, while its opponents tend towards an indeterministic *Weltanschaung.*[4]

Łukasiewicz seems to suggests that there is a connection between bivalence and metaphysical positions represented by the determinism/indeterminism controversy. However, this dependence requires a further analysis. For instance, we can ask what is prior, logic or determinism (indeterminism), that is, what provides premises and what constitutes the conclusion. Since the ancients were unclear at the point, Łukasiewicz cannot be blamed that his parenthetical remark is incorrect. His own reasoning, similarly as that of Waismann's, investigates the argument from bivalence to determinism. According to him (Łukasiewicz), bivalence and the principle of causality entail determinism. Is the principle of causality scientific or merely philosophical? Disregarding Łukasiewicz's own view, we can interpret his inference (logic plus causality $\Rightarrow$ determinism) either as based on scientific premises or mixed (one scientific, taken from logic and one philosophical). To complete this issue, let me note that most general as well concrete, systematic as well historical, elaborations looking at relations between philosophy and science consider both as co-existing and interrelated in many ways.[5]

A closer inspection of the relation between logic and determinism brings us to the next interpretative question. There are some minor differences between Łukasiewicz and Waismann. Whereas the latter speaks about the excluded middle and logical Predestination, the former refers to bivalence and determinism without further qualification. Yet we can overcome these disparities by saying, firstly, that Waismann employed the metalogical law of excluded middle, which functions as the most essential part of the principle of bivalence (in fact, the latter conjoins the former and the metalogical non-contradiction), and, secondly, pointing

[4] J. Łukasiewicz, "Philosophical Remarks on Many-Valued Logic", in J. Łukasiewicz, *Selected Works*, North-Holland 1970, p. 165 (this paper was originally published in German in 1931; tr. O. Wojtasiewicz)

[5] Here is a small sample of books discussing the relation between physics and philosophy: A. Eddington, *The Philosophy of Physical Science*, Cambridge University Press, Cambridge 1939, W. Heisenberg, *Philosophy and Physics*, Harper&Row, New York 1958, M. Čapek, *The Philosophical Impact of Contemporary* Physics, Van Nostrand, New York 1961, B. Gal-Or, *Cosmology, Physics and Philosophy. Recent Advances as a Core Curriculum Course*, Springer, New York 1981, J. T. Cushing, *Philosophical Concepts in Physics. The Historical relations between Philosophy and Scientific Theories*, Cambridge University Press, Cambridge 1998, R. Toretti, *The Philosophy of Physics*, Cambridge University Press, Cambridge 1999. For a comprehensive and up-to-dated survey, see *The Handbook of the Philosophy of Science*, ed. by D. Gabbay, P. Thagard and J. Woods, Elsevier, Amsterdam. The following volumes are available (I mention only titles and dates; particular books have own editors): *General Philosophy of Science: Focal Issues* (2006), *Philosophy of Logic* (2006),*Philosophy of Psychology and Cognitive Science* (2006), *Philosophy of Anthropology and Sociology* (2006), *Philosophy of Physics* (2007), *Philosophy of Biology* (2007), *Philosophy of Information* (2008), *Philosophy of Mathematics* (2009), *Philosophy of Technology and Engineering Sciences* (2009), *Philosophy of Statistics* (2011), *Philosophy of Medicine* (2011), *Philosophy of Complex Systems* (2011) and *Philosophy of Linguistics* (2012).

out that Łukasiewicz's determinism and Waismann's logical Predestination refer to the same philosophical position. However, other differences cannot be reconciled by so simple moves; Waismann explicitly says that he reconstruct Łukasiewicz's argument, but it is not quite true. As I already noticed, for Łukasiewicz, bivalence plus causality entails determinism, but Waismann's reconstruction omits causality. The crucial point is that Waismann denies that the (metalogical) excluded middle entails logical Predestination. He justifies his position to the use of "true" and "false" (details as irrelevant here). A lot of serious questions arises in this situation. Does Waismann's argument hold if we add causality to the excluded middle? What is the actual difference between both authors? Should we say that whereas Łukasiewicz argues that classical logic plus some additional premises imply determinism, Waismann says "since this argument is invalid for such and such reasons, classical logic does not entail determinism"? Łukasiewicz wanted to demonstrate that bivalence is incompatible with freedom and claimed that logic should be changed; he introduced many-valued logic for solving the problem. On the other hand, Waismann offered an argument for compatibility of logic and free action. I have no intention to decide who was right. My main task consists in showing how complex and many-sided is the application of logical theorems in order to derive from them philosophical statements.

We have to do with a fairly similar situation in the case of a famous controversy concerning the relation between quantum mechanics and determinism (and indeterminism, of course).[6] The most typical description is this (I omit the idea of hidden parameters advanced by Bohm and other proposals in the same spirit). Einstein and the representatives of the Copenhagen interpretations (Bohr, Heisenberg) appeared as the main protagonists. The former defended determinism, but Bohr and Heisenberg favored indeterminism. Einstein proposed various thought experiments, for example, that elaborated together with Podolsky and Rosen, in order to demonstrate that the Copenhagen interpretation was essentially incomplete. His opponents argued that Einstein's all attempts to abolish the "indeterministic" (I will later explain the use of quotes in this context) reading of quantum mechanics failed. Finally, Einstein agreed that since the Copenhagen interpretation is empirical faithful, he recognized it as legitimate, at least from the physical point of view. How to interpret this controversy? Did Einstein use the thesis of determinism as a premise in his arguments? Is so, his strategy is hardly comparable with that of Heisenberg who inferred the thesis of indeterminism from the

[6] See G. Auletta, *Foundations and Interpretation of Quantum Mechanics in the Light of a Historical Analysis of the Problems and of a Synthesis of the Results*, World Scientific, Singapore 2001 for a comprehensive survey. Of course, the scope this monograph (almost 1000 pp.) very considerably exceeds the determinism/ indeterminism/ quantum mechanics issue.

uncertainty principle, but not assumed the former in his reasoning. Should we say that Einstein rejected "indeterministic" consequences of the Copenhagen interpretation and thereby came to the conclusion that determinism was still tenable, but Heisenberg rejected determinism, because he deduced non-deterministic consequences from physics? Once again, we encounter here a very complex issue in which philosophical and empirical questions are mixed and interrelated in many ways. A striking fact is that natural scientists accepting the same empirical theories, share quite different, even inconsistent, philosophical views. This suggests that the premises/conclusion link without further clarifications does not suffice for accounting relations between science and philosophy. I will return to this issue after introducing a precise conceptual machinery. Looking at relevant texts, we encounter several other terms used in discussions about philosophical arguments based on science. Except "premise" and "conclusion", we have "supposition", "presupposition", "assumption", "consequence" or "result". I propose to consider the three first words as synonymous with "premise", but the two last as having the same meaning as "conclusion". I do not deny that there are other intuitions, for example, referring to subjective attitudes, styles of thought or even prejudices, but I tend to have devices subjected to logical analysis.

We have also to do with several accounts of the relation between premises and conclusions, like consequence of, entailment, derivation. following, implication or forcing. Let us agree that if **X** is a set of premises and *A* is a conclusion of **X**, we say that $A \in Cn\mathbf{X}$, that is, **X** is a logical consequence of **X** if and only if *A* can be formally derived from *X*. For simplicity, I equate the syntactic concept of logical consequence with the semantic concept of logical entailment (the set **X** entails *A* if and only if *A* is true in all models in which all sentences belonging to set **X** are true). Anyway, this description entails that rules of inference coded by *Cn* are infallible (correct, sound), that is, true premises inevitably lead to true conclusions. The metalogical characterization of the premise/conclusion relation forces a similar treatment of other methodological concepts. Let me list some definitions (they are simpliefed to some extent). The set **X** of sentences is a theory if and only if it is closed by *Cn* as an operation in the mathematical sense, that is, $Cn\mathbf{X} = \mathbf{X}$. Otherwise speaking, *X* is a theory if it is equal to the set of own logical consequences. Since the inclusion $\mathbf{X} \subseteq Cn\mathbf{X}$ is trivial (it directly follows from the definition of *Cn*), the substantial content of being a theory reduces itself to the inclusion $\mathbf{X} \subseteq Cn\mathbf{X}$. Thus, **X** is a theory if it contains own consequences. If there is a set $\mathbf{Y} \subseteq \mathbf{X}$ such that ,$Cn\mathbf{Y} = \mathbf{X}$, we say that **Y** axiomatizes **X** (**Y** is an axiomatic for **X**). Dependently whether **Y** is finite, infinite or recursive, we say that **Y** is finitely (infinitely, recursively) ax-

iomatizable. A theory **T** is consistent if and only if no pair $\{A, \neg A\}$ belongs to its consequences. **T** is (syntactically) complete if and only if for any A, $A \in Cn\mathbf{T}$ or $\neg A \in Cn\mathbf{T}$, and it is semantically complete if its every truth is provable from its axioms (one of my previous statements about *Cn* means that logic is semantically complete). Consistency is an obligatory property of theories (it practically means that inconsistent theories should be improved; this is common tendency in the history of science), but syntactic and semantic completeness are demanded, but, due to Gödel's theorems, inaccessible on level of arithmetic of natural numbers and beyond). If we take all arithmetical truth as axioms of arithmetic (of natural numbers), it becomes complete in both senses, although he is not finitely axiomatizable, because there are infinitely many true arithmetical assertions. However, and this is an important methodological observation, every theory is an axiomatic system.

The concept of theory in the metalogical (metamathematical) sense is an idealizations. In particular, any set of consequences of a given set of axioms is always infinite, but the actual theorizing is restricted to finite sets, because humans are able to effective cognitive acts operating on such collections. Hence we have a question how far the metalogical account of theories is faithful with respect to scientific practice. Since mathematics can be regarded as a collection of axiomatic systems, the metamathematical research widely exploits the concept of a theory as the logical closure of a given set of axioms. This perspective raises doubts as far as the matter concerns physics. Yet Hilbert in his famous lecture on mathematical problems delivered in 1900, raised the question of axiomatization of physics (problem 6), more precisely, he postulated a mathematical treatment of physical axioms, particularly of mechanics. Since he referred to earlier works of Mach, Boltzmann and Hertz, the issue was at stake about 1900. In fact, if **Z** includes Newton's three dynamical principles plus the law of gravitation, the set $\mathbf{T} = Cn\mathbf{Z}$ can be considered as an idealized picture of the classical mechanics. Further examples are provided by the relativity theory, quantum mechanics or quantum field theory.[7] Yet it would be difficult to maintain that axiomatic method became dominant in physics, even theoretical. On the other side, the following idealization is possible. We can consider even single physical laws together with their logical consequences as miniature theories. This is compatible with a notorious interest of physicists in particular theorems. Generally speaking, every theory **T** is formulated in a language $\mathbf{J}^{\mathbf{T}}$**.** We can identify **T** with a triple $<\mathbf{J}^{\mathbf{T}}, \mathbf{Y}, Cn>$,

[7] See H. Reichenbach, *Axiomatization of the Theory of Relativity*, University of California Press 1970 (German orinal appeared in 1924), G. Ludwig, *An Axiomatic Basis for Quantum Mechanics* I-II, Springer, Berlin 1985 and N. N. Bogolubov, A. A. Logunov, I. T. Todorov, *Introduction to Axiomatic Quantum Field Theory*, The Benjamin Cummins, London 1975. One can find further examples in J. Schröter, *Zur Meta-theorie der Physik*, de Gruyter, Berlin 1996 (this is a very comprehensive monograph about physical theories from the metalogical point of view) and G. Ludwig, G. Turler, *A New Foundation for Physical Theory*, Springer, Berlin 2006.

where **Y** is an axiomatic base, a collection of informal assumptions (postulates) or even a singleton. Although less mathematical fields, for example, chemistry and biology, are still less suitable to full and strict axiomatic reconstruction, but they fall under a more general model of theories, introduced above. I do not insist that single assertions with their logical consequences should be regarded as theories in the metamathematical sense, although I think that the triple $\langle \mathbf{J}^{\mathbf{T}}, \mathbf{Y}, Cn \rangle$ is an admissible approximation of $\mathbf{T} = Cn\mathbf{T}$.

The proposal to regard physical theories as axiomatic systems can be (in fact, it is the case) questioned by physicists. They will probably say that theories are rather models than set of sentences. I see no conflict here. We can consider theories as sets of sentences as well as speak about them as models. I would like also to stress that I do not claim that theories should be axiomatize or formalize. My enterprise is merely methodological and entirely belongs to philosophy of science. In particular, my special motivation consists in the decision to perform an analysis of the question undertaken in this study by the concept of logical consequence in its literal meaning. However, one can add something in favor of the "statement view of theories". First of all, physicists often say that theories are based on some postulates, for instance, that the velocity of light is constant. Secondly, they demonstrate something from the adopted postulates, for example that $v + c = c$, for every velocity v. These notorious facts allow to interpret postulates as axioms and demonstrations as proofs in the formal sense. Thirdly, physicists apply several metalogical concepts to physical theories, for instance, independence (of postulates), equivalence (of theories or postulates), extension or reduction (of theories) or consistency (of theories). Of course, one should be careful in using such analogies, because, for instance, Einstein's objection that quantum mechanics in the Copenhagen interpretation is incomplete did not refer to syntactic incompleteness, but pointed out that something was overlooked by Bohr and Heisenberg. However, such differences do not invalidate applying metalogic to analysis of empirical scientific theories. The skeptics with respect to the proposed analysis can eventually say that it does not produce so important results as it has place in metamathematics. I do not like to appeal to a typical answer that nothing should be decided a priori, although it is quite possible that investigations about computational complexity will find applications in physical calculations. I stress once again that my task is philosophical. I hope to show that treating physical theories as axiomatic systems allows to exhibit some misunderstandings concerning relations between science and philosophy.

If we adopt the proposed approach (even liberalized) to scientific theories, the problem of the relation between science and philosophy can be shaped in the following way. We ask whether philosophical statements occur among axioms of scientific theories and whether

philosophical assertions belong to *Cn***T**, where **T** is philosophy-free. One should remember that a given theory **T**, independently of its understanding as **T** = *Cn***T** or **T** = <$\mathbf{J}^{\mathbf{T}}$, **Y**, *Cn*>, contains its axioms and their consequences and nothing more. Thus, the set *Cn***T** forms the scope (or the domain of application) of **T**. Clearly, the scope of a given theory **T** determines its limits (borderlines) as well. This statement has a clear meaning only in the case of considering assumptions used in theories as axioms and conclusions derived from them as logical consequences. It is easily to confirm by numerous historical data that scientists *qua* scientists are not ready to extend the scope of theories by philosophical assertions. Let me illustrate this tendency by concrete examples. The thesis that every phenomenon is defined univocally by its mechanical parameters (position, mass, velocity) represents the core of the mechanistic world-view. Materialists of the 18th century supported this view by an appeal to classical dynamics (**CD**, for brevity). However, it was a considerable extension of the scope this theory (see also below). Its standard scope contains everything definable inside **CD** and nothing more. Even if philosophers find this formula as controversial and open for a further interpretation, the physicists have no doubt that the scope of **CD** and the scope of the philosophical mechanistic world-view are different. This precisely suggests that the mechanistic world-view is neither assumed (understood as an axiom) of **CD,** nor functions as its logical consequence. Similarly, philosophical atomism is neither an assumption of chemistry nor its consequence, and the same concerns the relation between vitalism and biology. Returning to determinism, indeterminism and physics, quantum mechanics neither assumes nor entails indeterminism, and classical physics has no inferential relations with determinism (see also below). This is the reason for writing "(in)deterministic" interpretation of quantum mechanics. To use a fashionable terminology, the language of physics (science in general) is incommensurable with the language of philosophy. This the main circumstance blocking the use of *Cn* across both.

My previous remarks does not imply that physics (or other science, but I concentrate on physics) has no connections with philosophy. The links between both appeared at the very beginning of European philosophy. The Ionian philosophy mostly concerned *physis*, that is, nature. Thus, philosophy and physics had the same subject and method in the first phase of philosophical thought and no matter if we will refer to theories of the Ionians as belonging to the philosophy of nature, physics or cosmology. In fact, they combined all these fields in the modern sense. It was Aristotle who explicitly distinguished *prote filosofia* (the first philosophy) as the science on being as being and physics as based on *empereia*. This distinction was respected by Archimedes and Ptolemy, two greatest ancient scientists, and never disappeared again. Newton's title *Philosophiae naturalis principia mathematica* does not provide any

counterexample, because it only witnesses a terminological custom of English nomenclature; speaking about physics as natural philosophy still occurs in British academic life. By the way, Newton's famous *hypotheses non fingo* can be understood as his claim that one should carefully distinguish philosophical hypotheses from statements based on experience. On the other hand, many physicists of the first rank, Newton himself, but also Galileo, Maxwell, Planck, Einstein, Bohr or Heisenberg, to mention only few, studied various philosophical problems. They considered them as important and published books titled *Physics and Philosophy* or somehow similarly (see note 5). Philosophers were (and still are) divided in their relation to physics as a source of philosophical insights. For example, Locke essentially used Newton's optical results, but Bergson or Heidegger maintained that physics has no importance for *the* philosophical understanding of reality. Disregarding thinkers unconditionally disrespecting the role of physics in philosophy, we encounter the question whether a physicist who discusses philosophical questions acts as just physicist or plays the role of a philosopher. In my view, he or she performs a philosophical job. Moreover, physicists *qua* physicists do not need to consider philosophical matters. This observations imply, contrary to Husserl (see the passage quoted from his *Ideas* above), that the link between physics and philosophy is factual, but not necessary. Yet I do not deny that philosophical views played (and still play) an important heuristic role in the development of physics. Einstein's belief that the world is well-ordered by the laws of nature motivated his attempts towards so-called deterministic interpretation of quantum mechanics. The reverse factual connection, that is, going from physics to philosophy should also be noted. For instance, the rise of classical mechanics resulted (philosophically) in the mentioned mechanistic world view, according to which h everything, including human action, is governed by the laws of dynamics. Since factual connections between physics and philosophy, although evident and frequently pointed out by historians of science (see Weiner's book mentioned in note 3), are not enough for accounting logical links between philosophical assertions and statements make by scientists *qua* scientists. Hence, my proposal to employ *Cn* and other metalogical tools in analysis of the main issue. The conclusion is negative: there are no logical relations between science and philosophy, provided that being a premise or conclusion is understood in the precise logical sense.

Yet the distinction between factual connections and logical consequence does not suffice for philosophical analysis of how science and philosophy are mutually related. To be more specific, the issue also concerns possible uses of scientific theories and assertions in philosophical debates. Let me return to some previously discussed questions in their typical traditional setting. I restrict my further remarks to so-called philosophical consequences of

physical theories. Once again, look at the mechanistic world-view as a consequence of **CD**, ask whether classical physics entails determinism and whether indeterminism can be derived from quantum mechanics. The sense of these (and similar) questions remains vague until we introduce references of "the mechanical world-view", "determinism" and "indeterminism" in a way acceptable for physicists and simultaneously compatible with philosophical intuitions, because this step means a necessary condition of using the phrases "consequence", "entails" and "can be derived" in the logical meaning. Otherwise speaking, we extend the intended scope of related physical theories by new phenomena. If, for example, the mechanical world-view is understood as the thesis that the entire reality consists of material points, which behave according to laws of **CD**, the extension in question appears as illegitimate until we show that, for instance, mental phenomena are mechanistic in this sense. Now the mechanistic conception of psyche is either correct or incorrect. If the first case occurs, the mechanistic world-view with respect to mental events becomes a trivial consequence of **CD**, but if the second alternative is adopted, this world-view must be qualified as an illegal extension of the scope of **CD**. However, the main philosophical problem consists in choosing one of possibilities occurring in the disjunction "the mechanistic conception of psyche is correct or incorrect" (and other similar dilemmas). For example, La Mettrie, the author of *Man A Machine* (the title is very instructive for materialism of the 18th century) was less interested in deriving his account of psyche from **CD** than in a materialistic analysis of mind. Thus, he choose the first member of the disjunction in question. Consider now the question whether quantum mechanics (**QM)** entails indeterminism. The uncertainty principle (**UP**) stating (I simplify) that $\Delta p_1 . \Delta p_2 \geq h$ (the uncertainty of position times the uncertainty of momentum is greater or equal to the Planck constant). This formula functions as the main premise of deriving indeterminism from quantum mechanics. However, it is problematic, because **UP** does not contains the word "indeterminism". According to elementary logic, a term occurring in a conclusion of deductive reasoning, must occur in its premises or be defined by earlier available linguistic means. Thus, one should introduce the term "indeterminism" (or "determinism") to **QM** in order to investigate the importance of this theory for the determinism/indeterminism issue. Heisenberg himself made such a step and said that determinism consists in predictability of future states of objects on the base of their past states. Since **UP** precludes the precise calculation of the past (including present) states, it also abolished the thesis of determinism on the level of the microworld and entails indeterminism. This argument is correct and shows how classical physics (**CD**, the relativity theory) differs from **QM**. Clearly, we are tempted to say that the latter is indeterministic, but to view the former as deterministic. Yet we have a variety of approaches

to determinism and determinism. For instance, the former can be defined not only *via* predictability, but also causally, statistically, probabilistically or by partial order in the Minkowski space. More importantly, different consequences of such definitions can be derived with respect to determinism and indeterminism of **QM**. Doubtless, all essential problems of **QM** and their solutions**,** can be formulated without any reference to determinism and indeterminism. On the other hand, what is important for philosophy does not directly follows from the literal content of physical laws. Incidentally, the same is to be said about so-called philosophical assumptions of science. In particular, they do not belong to assumptions made in deductions inside scientific theories.

However, some interpretative work is always done when philosophers use science in their arguments and speak, for example, that a physical theory has such and such philosophical consequences. The problem is that, on the one hand, we cannot handle this work as deriving philosophy form science, but, on the other hand, the reduction of the connection in question to merely factual coincidences seems not proper. What is going on? In my opinion, philosophers employ some hermeneutical operation (or insight), when they try to show that this or that scientific result has philosophical importance or not. This operation has in its background a postulate (the normative aspect of hermeneutic is substantial) that something, for example, determinism and indeterminism, should be understood in some way. One can look for hermeneutic hints in science, religion, ideology, politics, morality, ordinary life, etc., but I am particularly interested in hermeneutic insights motivated by science. If a hermeneutic is applied, further arguments can be deductive (this is frequent in the case of scientific hermeneutic), but they are mediated by an interpretation. Hence, we can label such consequences as interpretative. Briefly, indeterminism functions as an interpretative consequence of **QM**, modulo the definition referring to **UP**. The reason for adopting a hermeneutic are different. Doubtless, empirical data play a role in this respect, but they do not force solutions. Bohr's and Heisenberg's approach to philosophical problems of physics was definitely epistemological and motivated defining (in)determinism via predictability, but Einstein preferred the ontological way of thinking and believed in causality as the fundamental ingredient of determinism. Anyway, this perspective does not mean that science has no philosophical problems.

A few additional observations are in order. Firstly, every hermeneutic has its explicit roots in philosophical traditions. There is no other way of catching a given hermeneutic than embedding it into the history of philosophy, for instance, taking into account the development of the determinism/indeterminism debate. Secondly, there is no unique reading of data, in-

cluding theoretical and empirical ones, motivating hermeneutic interpretation. Thirdly, the adopted hermeneutic never liquidates a given philosophical controversy. Fourthly, explicit logical schemes of arguments supporting philosophical proposals are important, because they allow us to control arguments; hermeneutical parameters do not go against this function. Moreover, but it is related to my metaphilosophical view, the main philosophical aim does not consist in solving problems arising in philosophy, but rather making them explicit and clear. Thus, philosophical solutions are always relative to a given hermeneutic. Fifthly, the presence of hermeneutic in philosophy explains why philosophy basically remains in the same circle of problems and answers. However, there is no reason to be desperate by this fact. Every époque requires own philosophical hermeneutic, but it does not justify treating past hermeneutics as irrelevant. Although, as I earlier argued, indeterminism does not follows from **QM**, similarly as **CD** does not entail indeterminism, debating both philosophical views about the order of reality without taking into account modern physics, should be considered as irrational. On the other hand, there is probably no chance that philosophers ignoring physics in ontological or epistemological discussions disappear. This situation is regrettable for philosophers sharing my metaphilosophy, but should be tolerated. Sixthly, the role of philosophy in so-called context of discovery is obvious and cannot be denied. Even if we say that the borderline between discovery and justification is somehow vague, metaphysical views should not function as justifying scientific theories.

Let me finally consider the following view:

> It is, perhaps, easier to say what philosophy is not than what it is. The first thing, then I should like to say that philosophy, as it practiced today, is very unlike science; and this in three respects: in philosophy there are no proofs; there are no theorems; and there are no questions which can be decided, Yes or Not. In saying that there are no proofs I do not mean to say that there are no arguments. Arguments certainly there are, and first-rate philosophers are recognized by the originality of their arguments; only these not work in the sort of way they do in mathematics or in the science.[8]

Apparently, the view expressed by this quotation is puzzling. Waismann says that there are no proofs in philosophy, but they are arguments. We can add that philosophy has no deductive proofs, but deductive arguments occur in it. The problem is not verbal and cannot be answered by referring to the ambiguity of the word "proof". My reading of Waismann's diagnosis essentially employs the idea that hermeneutic parameter is substantially embedded in philosophical work; this concerns all kinds of philosophy, not only doctrines guided by methodological principles of analytic philosophy. The hermeneutical parameter just deter-

[8] F. Waismann, *ibidem*, p. 1,

mines that they are not proofs in philosophy, but arguments, deductive or not, related to hermeneutic. The latter are more or less original or even completely unoriginal, dependently of used hermeneutic and how it is done and developed. Philosophical considerations about physics belong to philosophy, not to physics even if they are made by physicists acting as philosophers. Otherwise, commensurability of science and philosophy can be achieved in philosophy via hermeneutic interpretation.

Ambivalence and Conflict: Catholic Church and Evolution[1]

gereon.wolters@uni-konstanz.de

I. Preliminary Conceptual Remarks

I would like to state one important point right at the outset. The Catholic Church has always maintained an almost enlightened position with respect to evolutionary theory, when one compares it with Christian American fundamentalism or its Turkish Islamic counterpart.[2]

There are, nonetheless conflicts. I would like to distinguish two types of conflict. The *first* is a doctrinal conflict in which science and religion hold conflicting, mutually exclusive, views about a particular situation. The most important example of this type of doctrinal conflict was seen in the case of Galileo and, to honour him, I term these kinds of conflict, *Galilean conflicts.* The most recent example of such a *Galilean conflict* is the debate surrounding evolutionary theory.

The *second* type of conflict is not so much about doctrine itself. It is more about scientists' attempts to refute that religion is a phenomenon in its own right. Such explanations are also called "naturalistic" or "scientistic". In this vein, Karl Marx described religion as the "opium of the people", Freud viewed religion as a collective neurosis and some modern brain researchers even regard it as an illusion produced by the limbic system. Others, in turn, see religion as an

[1] This paper derives from my *Ambivalenz und Konflikt. Katholische Kirche und Evolutionstheorie*, Konstanz (UVK) 2010, parts of which are included in my "The Epistemological Roots of Ecclesiastical Claims to Knowledge", in: *Axiomathes. An International Journal in Ontology and Cognitive Systems* (Dordrecht) 19.4 (2009), 481-508.

[2] During the meeting we got a vivid impression of the latter during a visit at Piri Reis school (Muğla) that is part of the Hizmet movement of the Turkish Imam and religious scholar Fethullah Gülen. According to Wikipedia the Gülen movement runs over 1000 Charter schools around the world, including 130 in the US. The schools are excellently equipped. Furthermore, education and science play an important role in Gülen's somewhat opaque teaching and even more opaque political practice. Nonetheless, biology was not mentioned when in a propaganda film mathematics and physics were praised. In private conversation one of the teachers referred to evolutionary theory as "monkey theory"… – The visit at Piri Reis was, by the way, requested as part of the funding by the Turkish Prime Minister's Promotion Fund.

important component of the evolution of social behaviour; while others like Richard Dawkins explain religion as a by-product of evolution. As in all these approaches religion appears as illusory, I would like to term these types of conflicts *Freudian conflicts*, because the word "illusion" appears in the title *The Future of an Illusion* of Freud's book on the topic.

II. Galilean Conflicts on Evolution

The Galileo affair has been a deep embarrassment to the Church ever since the second half of the 17th century when it became clear to almost everybody in Rome that Copernicanism was far from being "philosophically absurd and false" or "heretical".

Having become sort of prudent the **ecclesiastical authorities** kept a low profile throughout the first hundred years of Darwinian evolutionary theory. They seemed to have learnt their lesson from the Galileo Affair and kept their noses out of scientific debates, at least as far as making any official announcements about evolutionary theory is concerned. This is the more surprising as the topic of human evolution as – among other things – also dealing with the nature of man is much closer to central tenets of Faith than Copernicanism.

The first official and explicitly public and path breaking statement on evolution by a Church authority is the Encyclical *Humani Generis*, promulgated by Pope Pius XII in 1950. On the whole, this Encyclical expresses a rather relaxed position with respect to evolution. But although it does not instigate a Galilean conflict it nonetheless does intimate *possible* Galileo-like problems.

The text is somewhat obfuscated, however, by the low epistemological expertise, which has characterized documents of the Church up to the present day.

Talking about empirical science the Pope distinguishes between "clearly proved facts" and "hypotheses". However, as, by definition, all universal statements in empirical science are hypotheses, it seems more likely that the Pope is actually distinguishing between

hypotheses that are strongly supported by empirical evidence and hypotheses that lack sufficient empirical evidence.

In this light, we can say that Pope Pius XII:

1) accepts evolutionary theory as a scientific theory as long as it does not contest

a) God's creation of the human soul and b) the monogenic origin of mankind (which contradicts all scientific evidence)

2) The Pope requires that evolutionary "hypotheses" have to be "submitted to the judgement of the Church." Whether this also holds for "proved facts", remains unclear.

3) does not speak out on whether he thinks that evolution is a historical fact of the history of the earth.

The next pronouncement of the Church concerning evolution can be found in the *Monitum*, a warning against the writings of Jesuit palaeontologist Teilhard de Chardin, issued by the Holy Office on June 30, 1962 and reiterated on July 20, 1981.

The *Monitum* clearly illustrates two important points. 1) The Church is not interested in engaging in a Galilean conflict about evolution and explicitly refrains from interfering with matters of science. 2) The Church maintains a cautious and expectant position with respect to evolutionary theory.

This caution seems to be thrown to the wind in a famous letter by John Paul II to the Pontifical Academy on October 22, 1996. In this letter, Pope John Paul II confirms the position taken by Pius XII in *Humani Generis*, but with one decisive qualification:

"Today, almost half a century after the publication of the Encyclical [*Humani Generis*] new knowledge has led to the recognition of the theory of evolution as more than a hypothesis. It is indeed remarkable that this theory has been progressively accepted by researchers, following a series of discoveries in various fields of knowledge. The convergence, neither sought nor fabricated, of the results of work that was conducted independently is in itself a significant argument in favour of this theory."

But the Pope adds: “theories of evolution which, in accordance with the philosophies inspiring them, consider the mind as emerging from the forces of living matter, or as a mere epiphenomenon of this matter, are incompatible with the truth about man.”

1) Pope John Paul II acknowledges the theory of evolution to be an adequately confirmed theory or, as formulated in Vatican epistemological terminology, it has risen above mere “hypothesis” and is beginning to be something like a “proven fact”.

2) He, nonetheless, points to conflict areas. a) the monogenic origin of mankind (by implication only, because he confirms in a summary way what was said in *Humani Generis*) and b) God’s direct creation of the soul. The thesis of the monogenic origin contradicts scientific evidence about the formation of species, while the question of the soul is a special conceptual issue that, to the best of my knowledge, the pertinent sciences probably are not that concerned about. But that the Pope contests the evolution of mind and brain contradicts flatly his praise of evolutionary theory in general as well as well confirmed results of evolutionary theory, anthropology and palaeontology.

Given that general policy to get out of the Galilean fire line, it is most surprising that recently the Church, in the person of one of its most senior Cardinals, seems to have taken up arms again and marching head-long back on to this Galilean battlefield. In an article (“Finding Design in Nature”) that was published in the *New York Times* on July 7, 2005 Christoph Cardinal Schönborn was widely perceived as siding with the most recent incarnation of American Creationism, the so-called Intelligent Design Theory, ID for short. As this paper focuses on epistemological issues, I will not address all of the many other interesting aspects of this article but I will concentrate here on two pertinent quotations:

1) “The Catholic Church, while leaving to science many details about the history of life on earth, proclaims that the human intellect can readily and clearly discern purpose and design in the natural world, including the world of living things.”

2) “Evolution in the sense of common ancestry might be true, but evolution in the neo-Darwinian sense – an unguided, unplanned process of random variation and natural selection – is not. Any system of thought that denies or seeks to explain away the overwhelming evidence for design is ideology not science.”

As to the first quotation, I should remark that evolutionary biology in the course of its 150 years of existence has been able to explain thousands of design-like structures in living beings in terms of evolution, of which natural selection, as proposed by Darwin, is the most important but not the only factor. Before the advent of Evolutionary Theory, such structures were believed to have been drafted by an omnipotent designer. To answer Cardinal Schönborn’s first point in just one sentence: the human intellect, indeed, is able to discern purpose and design in the natural word, but explains this *scientifically* in terms of functional adaptations brought about mostly by natural selection.

As to the second point, tens of thousands of biologists all over the world will be astounded to hear that by relying on the two principles of evolutionary theory: random variation and natural selection, they are ideologists rather than scientists. Taking Cardinal Schönborn’s assessment seriously and dismissing random variation and natural selection would put an end to both evolutionary biology, and most other areas of biology, as we know them today.

Schönborn’s objections against evolutionary theory are, by the way, well known from creationist literature. Their mantra like repetition does not get them closer to the truth:

Both evolutionary biology as well as the philosophy of biology have dealt with these objections and have disproved them on countless occasions – to no avail.

Schönborn’s anti-evolutionism does not seem to be an isolated position, however. In September 2006 in Castel Gandolfo at a meeting on evolution of Pope Benedict with his former students, he praised Schönborn’s article in the *New York Times* this way.

„It occurs to me that it was divine providence that lead you, Eminency, to write a gloss in the *New York Times*, to render public again this topic and to show, where the questions are."

Normally, one finds even behind bizarre positions of the Church a rational core. This seems to hold also in this case:

It is not clear whether Cardinal Schönborn *really* intended to do what he actually did: launching a new Galilean conflict; and whether he really wanted to side with ID. There is some evidence that he did not want this and that he merely meant to engage in a *Freudian* conflict but that he applied the arguments the proponents of ID implement in their *Galilean* fight against evolutionary theory.

III. Freudian Conflicts on Evolution

Freudian conflicts arise, when a particular science tries to explain away religion as a phenomenon in its own right. They do not specifically affect the Catholic Church, but religion in general. Therefore, the first task of those who wish to wage a Freudian conflict should be to develop an adequate definition, or at least a satisfactory characterization, of the concept of religion. So far nobody seems to have achieved this and, unfortunately, most of those waging Freudian conflicts hardly even acknowledge this as a major problem. The second task would be to adduce sufficient scientific evidence in order to substantiate their Freudian claims in explaining religion.

These two defects one finds also in Richard Dawkins' *God Delusion*. In Chapter 5 ("The Roots of Religion"), it is clear that Dawkins has difficulties in pinpointing the direct adaptational value of religion. After rejecting explanations based on group selection, Dawkins starts with the confession: "I am one of an increasing number of biologists who see religion as a by-product of something else" (174). The idea of by-product, i.e. the idea that a structure that at some period in time had evolved according to certain selective pressures is later used for other purposes than the one it was originally selected for, is quite common in evolutionary biology. This phenomenon is called "exaptation" of a

structure, which is distinct from adaptation. Dawkins goes on to present the bold idea that: “natural selection builds child brains with a tendency to believe whatever their parents and tribal elders tell them. Such trusting obedience is valuable for survival.” (176). Religion is just a by-product of this brain structure.

Firstly, to assume that religion is above all or even exclusively about “trusting obedience” seems a rather narrow view of a monotheistic religion let alone a non-monotheistic religion. Secondly, as far as evidence is concerned, Dawkins just presents us nothing else than a just-so-story that abounds with “might” “could” and similar linguistic indicators of uncertainty and speculation. If natural science were conducted in this way, there could be no natural science in the sense we know and trust. In fact, Dawkins is much aware of the weakness of his position. “I must stress”, he admits “that it is only an example of the *kind* of thing I mean, and I shall come on to parallel suggestions made by others. I am much more wedded to the general principle that the question should be properly put [i.e. religion as a by-product of the evolutionary process], and if necessary rewritten, than I am to any particular answer.” (174). In response to this, it must be said that the very principle of scientific research is that ideas have to be supported by evidence. What is virtually missing from Dawkins’ claim is the evidence that religion is a “by-product of something else”.

My criticism of Freudian attacks on evolutionary explanations of religion given here has two targets: 1) I would like to contest their claims that they have scientifically explained away religion by means of natural science. At best they could show some behavioural dispositions for religion in humans that are far away from the phenomenal richness of religions. Generally we see here the problem of methodological naturalism: are the natural sciences the right way of dealing with cultural phenomena? My preliminary answer is: NO. Cultural phenomena are much too complex as to allow one-dimensional explanations.

2) But I would also like contest the claim that the self proclaimed “new atheists” have proven atheism to be true. Even if we concede, for arguments sake, that their evolutionary explanation of religion was correct, this would only show that humans have the corresponding behavioural dispositions (for social cohesion through religious symbols, obedience etc.). – A believer could easily answer that this only shows God’s wisdom in creation, insofar He/She has created us such that it is easy for us to believe in Him/Her.

MOORE'S PROBLEM

1. Moore's Original Versions of Moore's Problem

In 1942, Moore first presented the problem now known either as 'Moore's Problem' or as 'Moore's Paradox'. It was introduced by means of the following example:

(1) Although it may be true both that I went to the pictures last Tuesday and that today I don't believe that I did, it would be 'perfectly absurd' for me to assert the sentence 'I went to the pictures last Tuesday, but I don't believe that I did' (cf. "A Reply to my Critics" in *The Philosophy of G.E. Moore* (ed. by Schilpp, P.). La Salle (IL): Open Court, 1968, p. 543).

Later, in another essay, he used a different sentence, both in terms of content as well as structure, to build another example of the absurdity involved in its assertion. The sentence in question was the following:

(2) 'I believe he left, but he didn´t do it' (cf. "Russell's "Theory of Descriptions"" in *The Philosophy of Bertrand Russell (ed. by Schilpp, P.).* Evanston (IL): Northwestern, 1944, pp. 175-6.

2. The Oddity of the Absurdity

According to Moore's own view, the 'absurdities' he identified by means of the examples above were to be generalized to all cases in which we produce an assertion either of the form 'p and I don't believe that p' (as in (1) above), or of the form 'p and I believe that not p' (as in (2) above).

At the same time, Moore also pointed out that the identification of these absurdities cannot fail to strike a critical thinker as being somehow odd.

This oddity manifests itself in the fact that, as soon as we identify the assertions of the forms above as absurd, we are led to ask ourselves the following question: How can the assertion of a meaningful conjunctive sentence, the conjuncts of which may both be true simultaneously on many an occasion, be absurd?

In order to get a clear view on Moore's problem, we need to be able to find a plausible answer to this question; if we are not able to do that, then the oddity Moore detected is probably best seen as a symptom that something is not right with the original diagnosis.

3. The Most Travelled Route

In general, philosophers dealing with Moore's problem followed the route of assuming the intuition of absurdity associated with the actual use of sentences of the form 'p and I believe that not p' or 'p and I do not believe that p' to be legitimate (i.e., they assumed that the actual use of sentences of these forms really instantiates a paradox). They then proceeded from there in order to determine where the contradiction-like aspect of the problem that might justify such a diagnosis lied.

Their standpoints differed only in the different stories they presented in order to account for the emergence of this contradiction-like aspect. These stories admit being divided into two categories: those of a linguistic bent and those of a doxastic bent.

A. THE LINGUISTIC ANALYSIS

4. Moore's Own Way of Travelling the Most Travelled Route

Moore himself believed that the intuition of absurdity revealed the way in which assertion implies belief. His contention was that, whenever a speaker asserts that p, he also implies both that he believes that p and that he does not believe that not p.

This being the case, producing assertions of the above-mentioned forms would be absurd because what their second conjunct states *explicitly* contradicts what the first conjunct *implies*.

5. What is the Nature of the Implication? Moore's Reply.

As a matter of fact, the contradiction Moore claimed to have detected does not follow immediately from the implication he claimed there to obtain between assertion and belief. Besides the implication, some further assumptions need to be made in order for such a contradiction to be effectively derivable.

But, more importantly, to state that an implication exists linking the assertion of p with the belief that p and the absence of the belief that not p is not enough. The *nature* of such an implication must be clarified.

All the more so because, as Moore was the first to point out, such an implication is obviously not a matter of logical entailment. So, on what grounds are we to establish that a speaker who explicitly asserts p implicitly believes p and implicitly does not believe not p?

Moore's reply was that this implication is to be brought back to an *inductive inference*. According to him, we all learn from experience that, in the vast majority of cases, a man making an assertion believes what he asserts; i.e., lying, although possible, "is largely exceptional".

6. What is Wrong With Moore's Reply.

Moore's reply does not seem to be a correct analysis of the problem. If it were, the uttering by a speaker of an assertion of one of the two problematic forms mentioned above would be perceived by his interlocutors as clashing against nothing more than an expectation based on a previously observed empirical regularity.

However, the consequence of the perception of such a clash would probably be a reaction of surprise, followed or not, depending on the strength of the evidence, by a revision of the interlocutors' empirical expectations concerning the frequency of lying; hardly the conviction that they had witnessed the uttering of an absurdity.

Indeed, a genuine absurdity should result from a violation of a conceptual connection and not from a clash between the observation of an unexpected case and previously existing empirically based expectations.

7. The Wittgensteinian Analysis.

A more promising account of the nature of the connection between assertion and belief underlying the absurdity Moore detected comes from the Wittgensteinian tradition.

According to Wittgenstein's later philosophy, an important distinction in the deep grammar of ordinary language needs to be made between first-person singular present tense sentences with psychological content and third-person present tense sentences with psychological content. Whereas the latter are *descriptive* of the psychological reality of the person referred to by the personal pronoun, and thus susceptible of being true or false, the former are merely *expressive*; as such, they *vocalize* the psychological reality of the speaker; they don´t describe it. Vocalizations may be genuine or fake, but not true or false.

Thus, according to a number of philosophers belonging to this tradition (e.g. Malcolm, Heal or Linville and Ring), an assertion of the form 'I believe that p' merely *expresses* the speaker's belief that p; it doesn't *describe* it. The assertion of such a sentence by a speaker is then nothing but a semantically inert variant of the assertion by him of the sentence 'p'.

8. The Oddity Explained Away.

Under these circumstances, the truth-conditions associated with the assertions of 'p' and 'I believe that p' would be exactly the same. Both would be about the *world* and not about the speaker's *psychological life*. They would, namely, be about that segment of the world described by the proposition 'p'. And both would express (although not with the same force) the speaker's belief in the truth of the latter.

Therefore, an assertion of the form 'p and I don´t believe that p' or of the form 'p and I believe that not p' would be the assertion of a plain contradiction; in spite of the surface grammar of the propositions contained in them, both of these assertions would be of the form 'p and not p'. The intuition of absurdity Moore detected would thus be easily justified as a consequence of the underlying presence of a logical contradiction.

Thus, if the Wittgensteinian doctrine about the meaning of first-person singular present tense psychological sentences is to be accepted, no oddity associated with Moore's absurdity diagnosis will remain. The oddity will have been explained away.

9. What Happens When the First-Person Pronoun Is Used Referentially?

The main criticism the Wittgensteinian account invites us to make is that it contains an illegitimate generalization. That is, it is indeed true that there are cases in which an assertion of the form 'I believe that p' is used in the way the Wittgensteinian says it is; but there are also lots of other cases in which 'I believe that p' is used in order to *refer* to the fact of the speaker's believing in p and not merely to *express* the speaker's belief in p.

Assuming that such cases exist, as they clearly do seem to, how can we account for the absurdity Moore detected when one such sentence is conjoined in an assertion with the sentence 'not p'? Obviously, the Wittgensteinian solution, as it was expressed above, cannot tell us anything about these cases.

10. The Speech-Act Analysis .

An alternative both to the Moorean and the Wittgensteinian analyses of the nature of the implication is provided by the speech-act analysis (e.g. Burnyeat or Martinich). This analysis may be summarized through the following sequence of steps:

(1) It is constitutive of the speech-act of assertion that p that it be accompanied with the intention of providing the audience with information that p through their recognition that that is the speaker's intention.
(2) A speaker cannot be recognized by his audience to have the intention to provide them with the information that p, unless he is believed by them to believe that p. That is, a speaker's being believed by his audience to believe that p is *constitutive* of his being recognized by them as having the intention to provide them with the information that p.

(3) Therefore, from (1) and (2), it follows that it is constitutive of the enactment of a speech-act of assertion that p that the speaker strives to provide his audience with information that p by making himself believed by them to believe that p.

In other words, if a speech-act of assertion is performed, the audience should recognise that it is the speaker's intention that they should end up believing both the proposition that it is explicitly asserted by him and the proposition that he believes what he has asserted.

Now, given the analysis presented above concerning the nature of a speech-act of assertion, let's see what happens when the speaker asserts sentences of the form 'p and I don't believe that p' or sentences of the form 'p and I believe that not p'.

11. How Does the Absurdity Come About?

In the case of the assertion of a sentence of the form 'p and I don't believe that p', the propositions that the audience should recognise that it is the speaker's intention that they should believe are: 'p and I don't believe that p' and 'I believe that *p and I don't believe that p'*. Now, if we assume both the truth of the asserted sentence (remeber that, by itself, the sentence is consistent) and that the belief in a conjunction entails belief in each conjunct, an overt contradiction is derivable from them, namely, that the speaker believes that p and that he doesn't believe that p.

In the case of the assertion of a sentence of the form 'p and I believe that not p', the propositions that the audience should recognise that it is the speaker's intention that they should believe are: 'p and I believe that not p' and 'I believe that *p and I believe that not p*'. Assuming the truth of the asserted sentence and that the belief in a conjunction entails belief in each conjunct, although no overt contradiction is derivable from these propositions, an ascription to the speaker of two inconsistent beliefs is (namely, that he believes that p and that he believes that not p).

12. The Nature of the Absurdity.

In either case, the propositions that the audience should recognise that it is the speaker's intention that they should end up believing have consequences that clash whith each other. Thus, an audience guided by rational rules of conversational intercourse will be unable to make sense of the speaker's supposed assertion, given the fact that they will be unable to elicit from it any consistent intention of the speaker to make himself believed by his audience to believe the content of his own assertion.

This proof of how the inconsistency is produced is actually not the one the above-mentioned authors themselves present. But I think it is the right one. Anyway, and regardless of the details, this is why, according to this analysis, the production of Moore-like assertion-attempts is supposed to be self-defeating.

13. The Nature of the Implication and of its Violation.

Now, although the speech-act analysis agrees with the Moorean analysis to the effect that assertion implies belief, it disagrees with it regarding the nature of such an implication.

What the assertion of the problematic sentences violates, according to the speech-act analysis, is thus not an established empirical expectation but rather a set of conditions which are conceptually constitutive of the production of a legitimate speech-act of assertion. The absurdity is then the outcome of the speaker's use of the external indicators of the speech-act of assertion together with his violation of the internal conditions that constitute such an act.

The sentences of the form Moore identified are then deemed by the speech-act analysis to be *unassertable*, not in the sense that they cannot be uttered with an assertive tone of voice (which they obviously can), but in the sense that it is not possible to utter them and simultaneously fulfill the conditions that define the performance of a speech-act of assertion.

B. THE DOXASTIC ANALYSIS

14. Unbelievability.

More recently, a number of philosophers put forth the claim that it is misleading to view Moore's problem as having to do solely with linguistic expression (namely, with the violation of the conceptual conditions that are constitutive of the production of a particular kind of speech-act, viz., that of assertion). They feel that this diagnosis does not go deep enough.

They wish to make a stronger claim concerning Moore's problem, namely, the claim that contents of the form 'p and I don't believe that p' or of the form 'p and I believe that not p' are actually *unbelievable,* and not only *unassertable* (cf. Williams, Shoemaker, Sorensen). *A fortiori*, they wish to claim that it is *because* these contents are unbelievable that they are unassertable.

Thus, according to these philosophers, having the contents exemplifying Moore's problem as the objects of a propositional attitude such as belief violates conditions that are constitutive of meaningful *thought*. And this is why the intuition of absurdity is generated.

15. Logic as a Criterion of Doxastic Admissibility.

Their idea is then to replace with inner intrapersonal constraints of doxastic admissibility the interpersonal constraints that regulate, within the speech-act analysis, what is to count as an assertive move within the context of a theory of overt linguistic games.

But this is easier said than done. How are we to discover what these purely inner criteria of doxastic admissibility might be? This is a difficult problem brought about by this idea.

The view all these philosophers share regarding the nature of these intrapersonal constraints of doxastic admissibility is that they are of a logical nature. In particular, that it is the criterion of logical consistency that should do the job.

16. Inconsistency as a Criterion of Unbelievability.

Now, if logical consistency is the criterion in terms of which putative belief contents are to be assessed regarding their believability, then, if it is possible to show that a certain belief content is inconsistent or generates an inconsistency, then it has been shown that such a belief content is actually unbelievable.

Thus, the strategy followed by these authors in order to show that the propositions exemplifying Moore's problem are unbelievable is the strategy of showing that their admission as putative belief contents violates the criterion of logical consistency.

Contrary to Shoemaker's or Williams's, Sorensen's approach has the merit of not using in his proof of the unbelievability of contents exhibiting the forms Moore highlighted the principles B(p)→B(B(p)) (i.e., if the agent believes that p, then he believes that he believes that p) and B(B(p)) →B(p) (i.e., if the agent believes that he believes that p, then he believes that p). I deem this characteristic to be a merit of Sorensen's approach, because I take these principles to be highly contentious. They assume, namely, that belief is self-intimating. But this assumption seems to me to be plainly false. Thus, I think that Sorensen's views on this subject are those which best represent the standpoint I am now addressing.

Let us see then how his deductive strategy is supposed to work.

17. Proof of Inconsistency.

Let us consider first the case of my considering whether or not to accept a content of the form 'p and I don't believe that p' as the content of a putative belief of mine. If we assume both the truth of the proposition that defines this content, and the basic principle of doxastic logic according to which belief in a conjunction entails belief in each conjunct, then an overt contradiction is derivable from my putative belief in this true proposition, namely, that the proposition 'I believe that p and I don't believe that p' is true.

Let us consider next the case of my considering whether or not to accept a content of the form 'p and I believe that not p' as the content of a putative belief of mine. Again, if we assume both the truth of the proposition that defines this content and the basic principle of doxastic logic according to which belief in a conjunction entails belief in each conjunct, then, although no overt contradiction is derivable from my putative belief in this content, the holding by me of two strongly inconsistent beliefs is (namely, the holding by me of the belief that p and the holding by me of the belief that not p).

18. Blindspots for Belief.

Thus, in either case, if I am a rational and deductively competent believer, I will not accept holding belief contents as these.

As a matter of fact, if we assume logical consistency to be a criterion of belief admissibility, the conclusion to be drawn from the analysis displayed above must actually be stronger than the one that is expressed by the formulation contained in the previous paragraph. In fact, the conclusion must be that such contents are actually *unbelievable*, regardless of my idiosyncrasies as a believer.

According to Sorensen, the fact that we need to assume the truth of the propositions defining Moore's examples in order to derive their unbelievability, reveals that they mark out a particular type of propositions, namely, those he calls 'blindspots' of belief. According to him, Moore's main philosophical merit was twofold: the discovery that there are blindspots for belief and the discovery of what they are (cf. Sorensen 1988).

Moreover, the existence of such blindspots for belief is supposed to be a proof that the domain of the believable is only a proper subset of the domain of the true, and, therefore, that truth cannot be defined in terms of belief.

19. How About the Non-Obvious Cases?

But can logical consistency really be the standard by means of which we assess believability?

Bear in mind that a belief in a content of one of the forms Moore identified as problematic is not a contradictory belief *per se*. It is rather a belief from which either a contradiction or a strong inconsistency is *derivable*.

However, the contents of the forms Moore identified are not alone in being of this kind. For instance, there are contents from the belief in which a belief in a content of the forms Moore identified is derivable. Are those contents also unbelievable?

Consider the following two examples of such contents:

(1) The lottery paradox. The man who refuses to gamble believes of each lottery ticket that it is not a winner; however, he is aware that one of them will be a winner. Thus, he can be represented to believe a content that entails the content L such that L='W(1) or W(2) or ... W(n) and I believe that not W(1) and not W(2) and ...not W(n)'. L is, of course, of the form 'p and I believe that not p'.

(2) Sorensen's own atheism example. A more or less convoluted story can be concocted according to which it makes sense to imagine someone ending up believing the following content: 'The atheism of my mother's nieceless brother's only nephew angers God'. But belief in this content implies belief in the content 'My atheism angers God' which, in turn, implies 'God exists and I do not believe that God exists', which, of course, is of the form 'p and I do not believe that p'.

20. How Many Unbelievable Contents Are There Actually?

Now, I claim that although it may be epistemically wrong to believe in the truth of the contents above (as it certainly is), it is highly implausible to claim of them that they are unbelievable.

In order to strengthen my case, I ask you now to consider the case of other contents not related to Moore's propositional forms but that are also generators of inconsistencies.

As a classical example of one such case, consider the propositional content defining Axiom V of Frege's *Grundgestze der Arithmetik*. As Russell showed in 1902, this axiom generates a contradiction. But are we supposed to infer from Russell's proof that the Axiom V is actually unbelievable and that, therefore, between the 1880s and 1902, Frege was actually mistaken concerning his belief in the truth of Axiom V? I.e., that he only believed he believed in Axiom V but that, in reality, he didn't (because he couldn't)? This does not sound right.

21. Deductive Distance.

A possible way out of this conundrum might be to try to define a metric of deductive closeness and to identify within it a point separating small from large deductive distances. Thus, if a contradiction were deducible within a small deductive distance from a putative belief content, such a content would be unbelievable; if it took a long deductive distance to infer a contradiction from such a content, then it would be believable, despite the inconsistency it would lead to. This way we might get the means to distinguish in a rigorous way between acceptable and unacceptable forms of inconsistency.

The expectation would then be that the contents having the forms Moore highlighted would, according to this criterion, fall within the side of the barrier containing the unacceptable forms of inconsistency.

However, the very idea that there could be an absolute metric of deductive distance seems not to make much sense (cf. Cherniak 1986).

22. Sorensen's Way Out.

The idea of using logical consistency as a criterion of *empirical* belief ascription seems thus not to be very promising. Aware of this problem, Sorensen retreats to a normative standpoint according to which it is up to the rational observer to criticize the belief claims of the speaker. Such criticism is, in turn, to be developed in light of the desiderata of belief formation.

Sorensen's view is that *avoiding error* is the primary of these desiderata. And the structural constraint the fulfilment of which best serves it is logical consistency. This is therefore the criterion the following of which entitles us to criticize those who claim believing in contents from which inconsistencies are derivable and to urge them to revise their belief claims in order to eliminate the inconsistencies and preserve the consistency of their belief sets.

However, avoiding error and getting truth are not exactly congruent desiderata. This explains the existence of 'blindspots' for belief – true contents, the belief in which generates inconsistencies. The having of beliefs with these contents would violate the structural constraint put in place by the need to follow the primary desideratum of avoiding error. Therefore, such true contents cannot constitute any of our belief contents.

23. Actual Desiderata of Belief Formation.

But is it indeed sensible to imagine that the cognitive architecture of complex creatures should be best served by a mechanism of belief formation that strives first and foremost to avoid error?

I believe it is highly doubtful that this is so. Let me introduce what I take to be two counterexamples to this thesis.

Counterexample (1) is provided by the fact that living creatures in general (and not only humans) are prone to err on the side of caution. Arguably, this makes evolutionary sense. The following of rigorous processes of belief formation primarily aimed at avoiding error would, in many circumstances, simply be too costly and time-consuming. Presumably, for a whole range of creatures having to live, act and react quickly in the real world, the following of such a cognitive strategy would frequently be suicidal.

24. Usefulness and Truth.

Counterexample (2) is more parochial. It is provided by psychological research on belief in the hot hand in sports (cf. Burns 2001, 2004). Burns found out that belief in the hot hand is widespread among basketball players. He also found out that having this belief leads playmakers to pass the ball to a player with a higher scoring average in the game relative to his average performance and thus increases the chances of his team winning. The having of such belief seems thus to lead to the adoption of an adaptive behavioral strategy.

But 'hot hand' is defined as the higher probability in sports to score again after two or more hits compared with two or three misses; now, given the fact that each throw of the ball is actually independent of any other, belief in the hot hand is belief in a fallacy. The hot hand is, basically, an inverted version of the famous gambler's fallacy.

Thus, counterexample (1) pointed out circumstances in which following the cognitive strategy of avoiding error is presumably detrimental to the belief holder. And counterexample (2) above presented circumstances in which not following the cognitive strategy of avoiding error seems to be beneficial to the belief holders and the group to which they belong.

Taken together, counterexamples (1) and (2) suggest that the question of how useful a belief is in achieving some desirable goal should not be confused with the question of what its truth value is.

25. Cognitive Processes Ought to be Judged by Adaptive Criteria.

In reality, we simply don't know what are the general structural constraints for belief formation set by our cognitive architecture. This fact notwithstanding, counterexamples (1) and (2), and countless others in the literature (cf. Kahneman, Tversky, Gilovich, etc.) lead us to conclude that a structural constraint aimed primarily at avoiding error is not, both empirically as well as normatively, a serious contender for the job of determining belief admissibility. Cognitive processes, such as belief generation, ought to be judged by *adaptive criteria*. And adaptability is connected to the success of the actions beliefs do trigger in relevant contexts.

In fact, and *for good reasons* (namely, computational ones), it is likely that, for creatures like us, most processes of belief formation are of a fast, frugal and dirty nature and are responsive to localized structural constraints only (cf. Gigerenzer 2000). As a consequence, inconsistencies are to be expected to emerge within the belief system taken as a whole.

This being the case, it seems to be a bad move in cognitive thinking to assume, as Sorensen does, that the mechanism of belief formation of an autonomous living

system should ideally obey first and foremost the structural constraint of avoiding error, even if this is not the way things appear to have empirically evolved.

26. Conclusion.

The linguistic diagnosis of the Moorean absurdity is based on a plausible analysis of the conversational constraints underlying the rules that define an interpersonal linguistic game of information transfer and persuasion. Within such a game of persuasion, a move displaying a Moorean content seems indeed to be defying the rules that constitute it.

The doxastic diagnosis, however, does not seem to be able to pin down a plausible constraint in terms of which belief in referential contents of the forms Moore identified could actually be criticized as violating some constitutive condition of meaningful *thought*. Thus, I see no reason why such contents ought to be labelled as 'unbelievable'.

Finally, I would like to conclude by saying that, *as far as we now know,* and despite their potential for generating inconsistency, we cannot rule out on purely *a priori* grounds the possibility that true referential beliefs of the form 'p and I belief that not p' or 'p and I don't belief that p' may actually be usefully believed in a number of contexts.

Literature:

Almeida, C. de, "What Moore's Paradox Is About" in *Philosophy and Phenomenological Research,* Vol. LXII, 1, 2001, pp. 33-58.
Baldwin, T., *G. E. Moore.* London: Routledge, 1990.
Black, M., "Saying and Disbelieving" in *Analysis*, 13, 1952, pp. 28-31.
Burns, B. D., "The hot hand in basketball: fallacy or adaptive thinking?" in Moore, J.D. & Stenning, K. (eds.), *Proceedings of the Cognitive Science Society.* Hillsdale (NJ): Lawrence Erlbaum, 2001, pp. 152-157.
Burns, B. D., "Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand"" in *Cognitive Psychology,* 48, 2004, pp. 295-331.
Burnyeat, M. F., "Belief in Speech' in *Proceedings of the Aristotelian Society,* 1967-68.
Cherniak, C., *Minimal Rationality.* Cambridge (MA): The MIT Press, 1986.
Clark, M., *Paradoxes from A to Z.* London: Routledge, 2002.
Collins , A. W., "Moore's Paradox and Epistemic Risk" in *The Philosophical Quartely* 46, 1996, pp. 308-319.
Gigerenzer, G., *Adaptive Thinking – Rationality in the Real World.* Oxford: Oxford University Press, 2000.

Gilovich, T., Vallone, R., & Tversky, A., "The hot hand in basketball: On the misperception of random sequences" in *Cognitive Psychology,* 17, 1985, pp. 295-314.

Heal, J., "Moore's Paradox: A Wittgensteinian Approach" in *Mind*, 103, 1994, pp. 5-24.

Hintikka, J., *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Ithaca (NY): Cornell University Press, 1962.

Kahneman, D., *Thinking, Fast and Slow.* London: Penguin, 2011.

Linville, K. & Ring, M., "Moore's Paradox Revisited" in *Synthese* 87, 1990, pp. 295-309.

Malcolm, N., "Disentangling Moore's Paradox" in Malcolm, N., *Wittgensteinian Themes: Essays 1978-1989.* Ithaca (NY): Cornell University Press, 1995.

Martinich, J. P., "Conversational Maxims" in *Philosophical Quarterly,* 30, 1980.

Moore, G. E., "Moore's Paradox" in *G.E. Moore: Selected Writings* (ed. Baldwin, T.). London: Routledge, 1993.

Sainsbury, R. M., *Paradoxes*, Cambridge: Cambridge University Press, 1995.

Schilpp, P. A. (ed.), *The Philosophy of G.E. Moore.* La Salle (IL): Open Court, 1942.

Schilpp, P. A. (ed.), *The Philosophy of Bertrand Russell.* Evanston (IL): Northwestern, 1944.

Searle, J., *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press, 1969.

Shoemaker, S., "Moore's Paradox and Self-Knowledge" in Shoemaker, S., *The First Person Perspective and Other Essays.* Cambridge: Cambridge University Press, 1996.

Sorensen, R.A., *Blindspots.* Oxford: Oxford University Press, 1988.

Sorensen, R.A., "Moore's Paradox" in Dancy, J. & Sosa, E. (eds.), *A Companion to Epistemology.* Oxford: Blackwell, 1992.

Williams, J. N., "Moorean Absurdity and the Intentional 'Structure' of Assertion" in *Analysis* 54, 1994, pp. 160-66.

Williams, J. N., "Moorean Absurdities and the Nature of Assertion" in *Australasian Journal of Philosophy.* Vol. LXXIV, 1, 1996, pp. 135-49.

Wittgenstein, L. *Philosophische Untersuchungen.* Frankfurt am Main: Suhrkamp, 1977(1ª ed. 1953).

Wittgenstein, L., *Letzte Schriften über die Philosophie der Psychologie.* Frankfurt am Main: Suhrkamp 1989 (1ª ed. 1982).

Wittgenstein, L., *Vermischte Bemerkungen.* Frankfurt am Main: Suhrkamp, 1977.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

Newton on Matter and Space in *De gravitatione et aequipondio fluidorum*

Hylarie Kochiras

Abstract
This paper explicates the concepts of matter and space that Newton develops in *De gravitatione*. As I interpret Newton's account of created substances, bodies are constructed from qualities alone, as configured by God. Although regions of space and then "determined quantities of extension" appear to replace the Aristotelian substrate by functioning as property-bearers, they actually serve only as logical subjects. An implication of the interpretation I develop is that only space is extended by having parts outside parts; material bodies are spatially extended only in a derivative sense, via the presence of their constitutive qualities or powers in space.

Introduction

Newton develops his account of material body in what Howard Stein has called the "creation" story or hypothesis. This account has also been called the "determined quantities of extension hypothesis" (Slowik, 2009), since Newton marks the account as speculative and develops it by associating various conditions with "determined quantities of extension".[1] I shall follow Stein's terminology, however, for reasons concerning Newton's account of minds, as explained later.[2] Understanding the account of body depends upon properly understanding these determined quantities of extension and their relation to space (extension) itself. It is therefore important briefly to review *De gravitatione*'s claims about space.

Features of space

For Newton, space is an existence condition for any substance and "an affection of every kind of being".[3] This latter description refers to the manner of existing in nature, a manner of existing quite different from that of an abstract entity or a number, as J.E. McGuire has explained.[4] As

[1] See *De gravitatione* in *Isaac Newton: Philosophical Writings*, 27: "I am reluctant to say positively what the nature of bodies is, but I would rather describe a certain kind of being similar in every way to bodies..."; and 28: "And hence these beings will either be bodies, or very similar to bodies. If they are bodies, then we can define bodies as *determined quantities of extension which omnipresent God endows with certain conditions."*
[2] See Stein, "Newton's Metaphysics", 275. Slowik refers to that account of bodies as the "Determined Quantities of Extension" or "DQE" hypothesis (see "Newton's Metaphysics of Space", 2009, 438.) I follow Stein's terminology in part to avoid reifying the quantities of extension, and in part for a reason concerning minds, as discussed at the end of §4.
[3] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21.
[4] Pointing to the manuscript 'Tempus et Locus' (c. 1692-93), as providing "Newton's most succinct statement of how place and time relate to existing things". McGuire explicates that statement as follows: "Newton answers the question: what is it for

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

space is an affection of every kind of being, so is it a condition for their existence. As Newton asserts in a well known remark, one repudiating the concept of spirits as transcendent, "No being exists or can exist which is not related to space in some way. God is everywhere, created minds are somewhere, and body is in the space that it occupies; and whatever is neither everywhere nor anywhere does not exist."[5]

Since space is an existence condition of substances, it is not surprising that Newton takes it to have its own manner of existing. It is neither substance, he emphasizes, nor accident.[6] That it is not an accident inhering in a subject means, in part, that as an affection of every kind of being, it cannot be localized to any one being. Accordingly, it is independent of bodies; if all bodies were annihilated, it would continue to exist unchanged.[7] Space more nearly resembles a substance than an accident, Newton indicates, and as we shall see later, he ascribes a degree of "substantial reality" to it. Indeed, he cites it as the one thing that can in some circumstances be conceived apart from God—a feature he will use to attack Descartes' account of matter as atheistic.[8] Yet though it has some substantial reality, still space is not a substance. For one thing, it is "not absolute in itself, but is as it were an emanative effect of God."[9] Its not being absolute could not

---

anything to exist in nature? It is to exist in a place and at a time. As the text implies, existing in place and time is what counts as actually existing, in contrast, for example, to existing in the manner of an abstract entity or as a number. This contention is supported by Newton's use of the phrase 'rerum natura'....." (McGuire, "Existence, Actuality and Necessity: Newton on Space and time", 465)

[5] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 25.

[6] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21-22. The ultimate source of Newton's view that space is neither substance nor accident is Renaissance thinker Francesco Patrizi da Cherso (1529-1597). Patrizi additionally held space to be wholly distinct from body, indeed a condition for matter's existence, and to be immutable, indivisible, and immobile. See F. Patrizi, 'On Physical Space' (De Spacio Physico), translated and commentary by B. Brickman, Journal of the History of Ideas, 4:2 (1943), especially 224–245. As Edward Grant explains (*Much Ado about Nothing*, 206-207), Patrizi is also the source of a surprising explanatory remark following Newton's claim that space has distinguishable parts, whose common boundaries may be called surfaces. Newton then goes on to explain that in space there are "there are everywhere all kinds of figures, everywhere spheres, cubes, triangles, straight lines, everywhere circular, elliptical, parabolical, and all other kinds of figures, and those of all shapes and sizes, even though they are not disclosed to sight....so that what was formerly insensible in space now appears before the senses....We firmly believe the space was spherical before the sphere occupied it, so that it could contain the sphere....And so of other figures." (Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21-22).

[7] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 22. See also 21: as "an affection of every kind of being", it is not a "proper affection" which is to say an action.

[8] See Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 31: "If we say with Descartes that extension is body, do we not manifestly offer a path to atheism, both because extension is not created, but has existed eternally, and because we have an idea of it without any relation to God, and so in some circumstances it would be possible for us to conceive of extension while supposing God not to exist?" On space's inability to produce effects, see *Newton: Philosophical Writings*, p 21-22, 34.

[9] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 21. That space is not a substance cannot fully be explained by its dependence upon God, in virtue of being an emanative effect of God. For as will be emphasized later, Newton accepts not only the strong sense of substance but also the weak sense, which applies to things dependent upon God, in particular, created minds and bodies. Although I cannot here address the question of how Newton understands an emanative effect, I am sympathetic to McGuire's view that the relation of space to God is one of "ontic dependence". (See McGuire, "Existence, Actuality and Necessity: Newton on Space and time", 480: "the relation between the existence of being and that of space is not causal, but one of ontic dependence".) McGuire's view provides an alternative to the three that Gorham (September, 2011) identifies as 'Independence', 'Causation', and 'Assimilation'. Gorham defends Assimilation, arguing that space and time are attributes of God, and indeed identical to God (and thus to one another); see Gorham, September, 2011, especially 289-92 and 298-304.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

by itself explain why it is not a substance; for neither are created substances absolute in themselves, being dependent upon God. Yet created substances have a different relation to God, precisely in virtue of having been created. There is also another important difference. Substances act, whereas space produces no effects.[10]

Though neither substance nor attribute, space is not nothing, Newton emphasizes, for it has properties. The properties he describes indicate a Euclidean space, three-dimensional, homogeneous, and infinite. Space is also eternal and immutable, and though parts may be distinguished within it, those parts are motionless and indivisible.[11] It is these features—the immobility and indivisibility of space's distinguishable parts—that are especially significant for Newton's account of body.

The creation hypothesis and the definition of body

Newton develops his creation hypothesis in two stages, first ignoring mobility but subsequently introducing it. He begins from the realization that we can temporarily make regions of space impervious to other bodies by moving our own bodies into them, observing that this might somehow simulate the divine power of creation. By his will alone, God "can prevent a body from penetrating any space defined by certain limits".[12] Such an entity would either be a body, or would be indistinguishable from bodies by us.[13] For if God made some region above the earth impervious to bodies and all "impinging things", it would be like a mountain; it would reflect all impinging things, including light and air, and it therefore would be visible and colored, and would resonate if struck.[14]

These entities would be very similar to corporeal particles, Newton notes, except for this important feature: he has imagined them to be motionless. For an entity to be a body, or at least to resemble bodies in all humanly perceptible ways, it must be mobile. He therefore now adds

[10] As I argue in §4, Newton takes God to be identical to his attributes, and fundamental to his creative power, that is, omnipotence; yet in doing so Newton does not eliminate substance but rather gives a reductive account of it. I note here that I reject the interpretation recently advanced by Geoffrey Gorham, though his arguments are intriguing. According to Gorham, God is identical to his attributes, but his attributes include space and time, and hence he is identical to space and time. (See Gorham, September, 2011, especially 289-92 and 298-304). In §4, I indicate the difficulties I see with that view.
[11] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 22, 25, 26.
[12] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 27.
[13] Newton means to emphasize that we cannot know matter's "essential and metaphysical constitution" (*De Gravitatione,* in *Newton: Philosophical Writings*, 27), or indeed the essence of any substance. This conviction reappears in later writings, including the General Scholium, where he writes, "We certainly do not know what is the substance of any thing. We see only the shapes and colors of bodies, we hear only their sounds, we touch only their external surfaces….But there is no direct sense and there are no indirect reflected actions by which we know innermost substances."(*Principia*, 942.) In this respect his account of body is strongly empirical.
[14] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

that the hypothesized entities are capable of being moved from place to place, and in a law-governed way, a feature that is relatively new to conceptions of body.[15] Additionally, the entities can stimulate perceptions in minds and be operated upon by minds.[16] The hypothesized entities are now just like bodies, being perceptible, and having shape, tangibility, mobility, and the ability both to reflect and be reflected. They therefore could be "part of the structure of things", just like "any other corpuscle".[17] This enables Newton to provide a definition of body (insofar as we can know them).

> We can define bodies as *determined quantities of extension which omnipresent God endows with certain conditions*. These conditions are: (1) that they be mobile, and therefore I did not say that they are numerical parts of space which are absolutely immobile, but only definite quantities which may be transferred from space to space; (2) that two of this kind cannot coincide anywhere, that is, that they may be impenetrable, and hence that oppositions obstruct their mutual motions and they are reflected in accord with certain laws; (3) that they can excite various perceptions of the senses and the imagination in created minds, and conversely be moved by them, which is not surprising since the description of their origin is founded on this.[18]

One of the interesting things about this definition is that Newton sees it as serving theological goals, as will become evident from his commentary, and yet it is firmly rooted in experience. The fundamental features of our experience with bodies appear in the definition: their mobility; the mutual impenetrability that results in law-governed reflections of other bodies, light, and air; and the sensations they produce in us, such as those of color. Newton's remark at the end of the passage highlights the fact that experiences, specifically perceptions, make his description of the bodies' origin possible. For if bodies lacked the power to produce sensations, we could never have any ideas of them.[19] It is notable that Newton specifies condition (3), the power to produce

[15] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28. In an otherwise quite different thought experiment, which appears in *Le Monde*, Descartes imagines bodies that move "in accordance with the ordinary laws of nature"; see CSM 1, 90. Of interest here is Katherine Brading's article "On Composite Systems: Descartes, Newton, and the Law-Constitutive Approach" (2011).

[16] "For it is certain that God can stimulate, our perception by means of his own will, and thence apply such power to the effects of his will." (Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28)

[17] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

[18] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28-29. A definition given in 1678 by Robert Hooke contains some intriguing similarities. After asserting that the universe consists in body and motion, he writes, "I do therefore define a sensible Body to be a determinate Space or Extension defended from being penetrated by another, by a power from within." He also speculates that body and motion might ultimately be "one and the same". See Hooke, Lectures *Potentiae Restitutiva*, or of Spring, Explaining the Power of Springing Bodies, 1678, 338-340. How near the similarity really is, however, is a question I will not pursue here.

[19] Geoffrey Gorham interprets this remark very differently. On his view, Newton's remark that the description of bodies' origin is founded upon sensations indicates that he takes the capacity to produce sensations to be both necessary and sufficient for body-hood. In connection with that claim, Gorham argues that Newton ultimately sees his conditions of mobility and impenetrability as superfluous; these "do no independent work of their own"(Gorham, Jan.2011, 24). I contest Gorham's conclusion about those conditions in §2.5.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

sensations, as distinct from condition (2), impenetrability. One reason for distinguishing them is that in the hypothesis' context, the first creation of matter, impenetrability could not be sufficient to produce sensations in minds. For if any minds existed when God first created matter, no human bodies would exist to touch it, and so the mutual impenetrability of bodies could not then produce sensations in minds. Yet there is another explanation for including condition (3) as independent of condition (2): even in the context of actual experiences, Newton does not seem to consider sensations as explicable solely in terms of impenetrability. He rather seems to share a belief common in the early modern period—that while the contact of light particles with the eye and food particles with the tongue seem to play some necessary role, they are not sufficient for the production of sensation, and so some role must be attributed to God.[20]

The definition's third condition is thus the basis for Newton's claim that Descartes' account of matter leads to atheism, while his own confirms God's existence. As indicated above, he takes space to be the one thing sometimes conceivable apart from God, since it produces no sensations or other effects, and so by identifying matter with extension (space), Descartes allows that matter is conceivable apart from God.[21] For as Newton indicates elsewhere, "we find almost no other reason for atheism than this notion of bodies having, as it were, a complete, absolute and independent reality in themselves."[22] On his own account, bodies are not conceivable apart from God, because their capacity to produce sensation cannot be so conceived, and that inconceivability is expressed directly by his definition's third condition.

---

[20] Here I disagree with Geoffrey Gorham, who argues that Newton actually intends his third condition, the capacity to produce sensations in minds, to resolve a problem about distinguishability (a problem that has concerned several commentators but did not, in my view, concern Newton, for reasons I indicate later in this section). On Gorham's view, if Newton did not intend his third condition to resolve that problem, it would be superfluous: "If the DQE's are impenetrable, they will be solid to touch, reflect light, perturb the air when struck, and so on. Since these are the means by which the senses perceive familiar bodies, why the need for God to affix also the special power to produce sensations? The answer seems to be that impenetrability alone is inadequate to distinguish bodies from the unfavored portions of absolute space." (Gorham, January 2011, 23). Yet as I have argued, Newton does not see the production of sensation as reducible to impenetrability, either in the context of matter's first creation, when no human bodies would exist even if minds did, or in his actual context, in which human bodies do exist. He takes a line similar to that found in Locke's *Essay*. Despairing of the ability of the mechanical hypothesis to reduce sensations to the shapes, sizes, and motions of particles, Locke suggests that the production of sensations must be attributed to God. Or, on an interpretation associated with Ayers, Locke thinks that we invoke superaddition because our powers of understanding are too limited to grasp how God might have enabled matter to produce sensations; my thanks to James Hill for discussion of the point.

[21] "If we say with Descartes that extension is body, do we not manifestly offer a path to atheism, both because extension is not created but has existed eternally, and because we have an idea of it without any relation to God, and so in some circumstances it would be possible for us to conceive of extension while supposing God not to exist?" (*De Gravitatione*, *Philosophical Writings*, 31). Interestingly, Newton's language here suggests the strong mental exercise that Descartes calls 'exclusion', as opposed to the weaker one of abstraction. For Descartes, a successful attempt to conceive something while actually separating or excluding another reveals that the two are really distinct, as opposed to being merely conceptually distinct but really identical; see Pr I.62, CSM, 214. Newton's phrase, "supposing God not to exist", suggests the strong mental act of exclusion; he suggests that space may be conceived while actually excluding God, by supposing him not to exist.

[22] *De Gravitatione*, in *Philosophical Writings*, 32.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

Interpreting Newton's account: determined quantities of extension and the role of divine action

Yet what exactly are the "determined quantities of extension" endowed with the three conditions that Newton asserts? The question is essential to an understanding of his account of body, but it also has implications for the nature and extent of divine providence, as we will see. It is often supposed that in his creation hypothesis, Newton takes God to create bodies from parts of absolute space itself. For example, Christopher Conn speaks of a body in *De gravitatione* as "nothing more than a divinely-modified region of space".[23] Geoffrey Gorham also takes Newton's determined quantities of extension to be parts of absolute space itself, contrasting the "favored regions of space", which God endows with powers, against the "normal" regions (though on his soft occasionalist interpretation, the favored regions of space are given only powers of producing sensations.)[24] If Newton were seeking some sort of substrate in which properties could inhere, space might initially seem suitable, since as noted earlier, he considers it to be more like a substance than an accident. Nevertheless, there are powerful reasons to deny that he supposes God to create bodies by modifying parts of absolute space itself.[25]

---

[23] Conn, 1999, 316, n. 23. Alan Gabbey allows the possibility without committing to it, in the following passage: "But alternatively, and of equal possibility, the properties of bodies might be the result of God choosing to 'inform' extensions, parts of absolute space, with corporeality and mobility. The parts of absolute space that God can and perhaps does endow with the properties of bodies are as empty of matter as the *materia prima* of the scholastics is void of intelligibility, or bereft of existence. But there is a crucial difference. Each of these parcels of empty extension is a *quid,* and a *quale,* and a *quantum,* whereas *materia prima* is none of these." (Gabbey, "The term *materia* in Newton and the Newtonian Tradition", 438). I implied this myself in an earlier article (Kochiras, 2009, 269).

[24] See Gorham, "How Newton Solved the Mind-Body Problem", January, 2011, 22: "Newton proposes that God creates bodies by imposing three conditions on certain regions of space or 'determinate quantities of extension'(DQE).") See also Gorham, "Newton on God's Relation to Space and Time: The Cartesian Framework", September, 2011, esp. 297, where he speaks of "a favored portion of extension".

As a result of taking this line, Gorham understands Newton's account of body as intended to respond to a problem of distinguishing the favored regions of space from the normal ones. The problem (a variant of which was raised by Bennett and Remnant, 1978), may be described by the following two claims. (i) Newton claims that the parts of space are immobile, and therefore the favored portions of space must be distinguishable from the normal parts of space in order to become mobile; yet (ii) the property of impenetrability cannot accomplish the task of making the favored portions of space distinguishable from the normal parts of space, because the normal parts of space are themselves impenetrable to one another precisely because they are immobile. This problem, and the need to resolve it, then motivates Gorham's interpretation of Newton's account of body. In Gorham's view, Newton intends the third condition of his account, i.e., the capacity to produce sensations, to resolve the problem, for in his view, that condition would be superfluous if not intended for that purpose. (Gorham writes, "Condition (3) solves this problem by ensuring that the favored regions of space stand out because God superadds to them something lacking from the unfavored regions: the power to produce sensations." Gorham, January, 2011, 23.)

But the third condition would not be superfluous absent that problem, as I argue in §2.5. Nor is it clear that the problem about distinguishability, which motivates Gorham's account, is genuine. For one thing, if God did modify parts of actual space, surely he himself could distinguish them from one another (as indeed he would have to be able to do, if he were to confer any properties at all upon them.) For another thing, as I argue, Newton's creation story and its associated definition of body does not suppose parts of space itself to be modified. And there is an even more important consideration: even if the problem were genuine, why should we allow the need to resolve it to color our interpretation of Newton's account, given that he himself is not addressing such a problem? Even if the problem were genuine, it should be invoked only to evaluate Newton's account, not to interpret it, since again, Newton himself is not addressing that problem.

[25] It should be noted that despite taking parts of space itself to figure in Newton's account of body, Gorham ultimately defends a soft occasionalist interpretation, on which Newton takes the regions of space to be modified only to the extent of temporarily

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

The starting point of the creation hypothesis, though hardly decisive, is potentially significant. That starting point is the observation that we can make spaces impenetrable by moving our bodies into them—an action that does not, notably, alter the nature of space itself. Also significant, I think, is the "metaphysical truth" that God "has created bodies in empty space out of nothing"[26]; to square his account with that truth, as he means to do, Newton cannot say that God creates bodies out of space, since space is not nothing. A consideration that should be decisive, however, is the nature of space as he describes it, together with the implications of supposing that actual parts of space figure in his creation story and definition. He described space as being eternal, immutable, immobile, unable to produce effects, and as having parts that are distinguishable but indivisible. To suppose that certain parts of space could be divinely modified, rendered able to produce sensations, solidified and set into motion, is to suppose a full contradiction of Newton's claims . It is to suppose that space is not eternal, because some parts of it may be turned into bodies; that space is not immutable, because some parts could be made impenetrable and able to produce sensations; and that its parts are not immobile and indivisible, because some parts, once made impenetrable, could be torn away from their neighbors and set into motion. And if some parts could be torn away, what exactly would ensue—would space be left with gaps, or would additional space appear to fill the gaps?

These are the sorts of conceptual problems that Newton points to when clarifying the first condition of his definition. Mobility is the first stated condition with which determined quantities of extension are endowed, and since space is immobile, he immediately clarifies that he is not speaking about the parts of space itself, but rather about their quantities: "therefore I did not say that they are numerical parts of space which are absolutely immobile, but only definite quantities which may be transferred from space to space."[27] Significantly, a quantity of some part of space is not identical to the part of space itself—after all, some numerically distinct parts of space have the same volume. Thus as Newton's own clarification indicates (a clarification we should keep firmly in mind when he seems to stray from it by employing more abbreviated

---

assuming powers to produce sensations in minds. For as noted in §2.5, Gorham argues that the first two conditions of Newton's definition turn out to be superfluous, and the "favored" parts of space, instead of being made actually impenetrable and actually torn away from the "normal" regions of space, are simply "spatial occasions" for God to produce perceptions in minds. Denying that Newton takes the parts of space to be altered and torn apart seems especially important for Gorham since he also argues that space is ultimately identical to God. Therefore, allowing that space could be altered would not only conflict with Newton's claim that space is immutable, it would also imply that God is not immutable; Gorham avoids that implication by arguing that conditions (1) and (2) of the definition "do no independent work".

[26] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 31.

[27] Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 28.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

locutions[28]), it is a mistake to reify his determined quantities of extension, by mistaking them for parts of space itself.[29]

Since Newton associates only quantities with the qualities or powers identified by his three conditions, and not parts of absolute space itself, bodies are constructed from powers alone. Insofar as it is useful to speak in terms of subject and the properties predicated of it, the quantity of any given region of space in which the powers are present may serve as a logical (grammatical) subject, but the utility of such locutions should not lead us to suppose that bodies consist in anything beyond powers. There is nothing like a substrate. Rather, bodies consist in sets of powers, distributed at multiple points of one region of space if the body is resting, or at points of successive regions if the body is moving. This interpretation does require that Newton's first condition, mobility, be considered differently from the other two, in that mobility must apply to something. I therefore suggest that Newton takes bodies (insofar as we can know them) to consist in mobile sets of spatially configured powers for mutual impenetrability and production of sensation. These mobile sets of powers must somehow be unified, so as to maintain their characteristic configurations as they either rest or move through space, and I propose that he assigns the task of unifying them to God. The powers are unified and maintained as enduring configurations by God—by y^{e} divine arm, to borrow a phrase that Newton uses elsewhere.[30] The divine will accomplishes the task that he takes to be performed in the Aristotelian account by prime matter or substrate.

This interpretation fits well with his emphasis upon perceived qualities as the basis of a substance. In one of the explanatory points following his definition of body, he explains that the entities he has described are no less real than bodies and may be called substances bccause

[28] At one point, for instance, Newton speaks of the form that God "imparts to space".(*De gravitatione*, in *Newton: Philosophical Writings* 29) Because of such instances, commentators must choose between (i) accepting the surface meaning of such remarks and thus understanding bodies as mobile, solidified regions of space, while paying the price of implying a serious conceptual problem (the question of what would remain, if regions of space could be torn out) as well as conflicts with Newton's own claims (i.e., that space is immutable and immobile, and that his definition concerns definite quantities, not the numerical parts of space); and (ii) avoiding any conflict with his claims that space is immutable and immobile, while paying the price of implying that some of his locutions are abbreviated or careless. I argue for the latter option, as indicated throughout.

[29] My interpretation can be reconciled with the definition that Newton gives of body at the outset of *De gravitatione* (and I thank Eric Schliesser for reminding me, at the conference at Ghent, of the need to reconcile them). As is well known, the bulk of *De gravitatione* consists in a lengthy digression, in which Newton attacks Cartesian physics and addresses various metaphysical questions, including those focused upon here. But Newton begins the manuscript with the intention of treating the weight and equilibrium of fluids and of solids in fluids, and while still engaged in that project, he defines body as "that which fills place" (*De gravitatione*, in *Newton: Philosophical Writings*, 13.) On the interpretation that I develop, that definition can be retained, since a set of spatially distributed powers of mutual impenetrability will repel any other such set; and while such sets do not fill place by actually having parts outside parts, the phenomenal effect is the same.

[30] The phrase is from Newton's second letter to Bentley (17 January, 1692/93; 240 in Turnbull): "Secondly I do not know any power in nature wch could cause this transverse motion without ye divine arm."

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

"whatever reality we believe to be present in bodies is conferred on account of their phenomena and sensible qualities."[31] And a remark elsewhere in the manuscript, which I discuss in more detail in a subsequent section, points to attributes as the basis of "substantial reality". An interesting implication of my interpretation is that the extension of bodies is parasitic upon the extension of space. Since bodies are extended in virtue of the presence of their constituent qualities or powers in space—a view whose conceptual predecessor is a concept of immaterial spirits as spatially located powers, as noted later[32]—only space is extended in the sense of having parts outside parts, a complete reversal of the Aristotelian view that all extension is corporeal, an attribute of matter.

The account of body and the extent of God's providence

In another of the explanatory remarks following the definition of body, Newton states that the entities he has described subsist "through God alone".[33] The interpretation I have given provides a specific way of understanding this: the entities subsist through God alone in that the sets of powers are unified and maintained in their configurations by divine action. Since this action is direct, God's providence is much greater than if he merely concurred with the bodies' continued existence. Still, Newton also leaves ample room for secondary causation, for as indicated earlier, he sees the account of body and thus God's direct action as limited to corpuscles. This suggests a view similar to that found in a much later text, Query 31 of the *Opticks*. Query 31 sidesteps the problem of cohesion at the sub-corpuscular level by suggesting that corpuscles are created by God, but it speculatively attributes the cohesion of aggregate bodies to interparticulate forces, and thus to secondary causes.[34] Here too, by restricting his account of bodies to corpuscles, Newton leaves the cohesion of aggregate bodies to secondary causes.

The role that Newton assigns to God in *De gravitatione* therefore falls considerably short of occasionalism. This is consistent with the expectations that he evinces in other texts. In a letter of 1680, Newton writes, "Where natural causes are at hand God uses them as instruments in his works".[35] And as I have argued elsewhere, Newton never endorses the hypothesis that God

[31] This claim appears in the second of the four explanatory remarks following Newton's definition of body; *De gravitatione*, in *Newton: Philosophical Writings*, 29.
[32] For a discussion of concepts of spirits and space, see Kochiras, 2012.
[33] *Ibid.* Newton, *De gravitatione*, in *Newton: Philosophical Writings*, 29.
[34] An illuminating discussion of Locke and the foundational problem about cohesion may be found in James Hill (2004), "Locke's Account of Cohesion and its Philosophical Significance".
[35] Newton to Burnet, 1680; Newton, *The Correspondence*, II, 334.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

causes gravitational effects directly, and his ongoing search for an explanation expresses his expectation of secondary causes.[36]

I therefore disagree with the interpretation defended recently by Gorham, who attributes occasionalism to Newton, albeit a soft sort.[37] The occasionalism is soft in that God does not cause perceptions in minds directly, instead endowing varying regions of space with the power to do so, in a continuous creation of matter.[38] Yet it is still a kind of occasionalism, because Gorham argues that that the first and second conditions of Newton's definition of body are superfluous, doing "no independent work of their own",[39] and that bodies consist in only the powers to produce sensations. Regions of space are the "spatial occasions" for the sensations, and God creates matter continuously by creating the powers to produce sensations in varying regions of space.[40] Gorham claims a powerful advantage for his interpretation: it implies that Newton solves the mind-body problem, avoiding problems about mental causation "by embracing a quasi-idealistic ontology of matter."[41] Yet his interpretation requires us not only to accept that conditions (1) and (2) of Newton's definition are superfluous, but also that condition (3), the power to produce perceptions in minds, is not merely necessary for body-hood but also sufficient. Gorham reaches this latter conclusion partly through his reading of the comment that Newton adds to this third condition—that it is not surprising that bodies have the power to cause perceptions in minds, "since the description of their origin is founded on this".[42] Yet there is a

---

[36] See Kochiras, 2009, 2011.

[37] Gorham indicates that he sees Newton as belonging to a tradition that locates the ground of causation in God's will (Gorham, January, 2011, 25).

[38] See Gorham, January, 2011, 24.

[39] Gorham, January, 2011, 24.

[40] See Gorham, January, 2011: "The continuous creation of matter amounts simply to the distribution within space of God's power to produce sensations"(24); and "various quantities of extension are the mere 'spatial occasions' for God to bring out our perceptions in the successive and law-like ways we associate with moving bodies."(25).

[41] Gorham, January 2011, 30.

[42] *De gravitatione*, 29. There is another passage that Gorham interprets as showing that Newton takes condition (3) to be sufficient as well as necessary for being a body. In that passage, Newton is attacking the Cartesian view of matter:

> "Let us abstract from body (as he demands) gravity, hardness, and all sensible qualities, so that nothing remains except what pertains to its essence. Will extension alone then remain? By no means. For we may also reject that faculty or power by which they [the qualities] stimulate the perceptions of thinking things. For since there is so great a distinction between the ideas of thought and of extension that it is not obvious that there is any basis of connection or relation [between them], except that which is caused by divine power, *the above capacity of bodies can be rejected while preserving extension, but not while preserving their corporeal nature*."(Newton, *De gravitatione* in *Isaac Newton: Philosophical Writings*, 33-34; emphasis added)

Commenting upon this passage, and quoting the italicized portion, Gorham writes, "So, the capacity to produce sensations in minds is sufficient and necessary for a quantity of space to possess the nature of body. This explains why Newton privileges condition (3) when he introduces his theory of creation: "The description of their [bodies'] origin is founded on this" (*De Grav* 29)."(Gorham, January, 2011, 24.) I do not see how Newton's remarks imply that condition (3) is sufficient as well as necessary for body-hood, as Gorham takes it to do. There is certainly a way of understanding the passages that does not imply any such thing. As I read the passage, Newton is saying that if one mentally abstracts qualities such as hardness away from a body, one

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

natural reading of that remark which does not require either dismissing the definition's first two conditions as superfluous or supposing the third to be sufficient. That natural reading, which I explained earlier, is simply that if bodies lacked the power to produce sensations, we could never have any ideas of them. The remark is an instance of Newton's oft-repeated acknowledgement that we can know only perceived qualities, not the "essential and metaphysical constitution" of things.[43] Since I reject the occasionalist interpretation, I also reject Gorham's conclusion that "Newtonian bodies do not seem to qualify as self-standing substances".[44] On my interpretation, Newton considers bodies to be created substances. This is a desirable result, since bodies would have to be substances in order for Newton to accept a substantial distinction between mind and body—and he does, as I argue elsewhere.

In closing, I suggest that the account of body Newton develops in *De gravitatione* might have indirectly helped facilitate a concept belonging to his later rational mechanics, that of point mass. On the interpretation I have given, his concept of body has as its conceptual ancestor a spirit which consists in causal powers, which lacks parts outside parts, and which is extended only in the derivative sense that its constituent causal powers are present in some extension. An entity consisting in spatially present causal powers, as opposed to one possessing parts outside parts, may more easily be conceived as existing in a larger or smaller area—even as contracted to a point. Thus the bodies of *De gravitatione*, which consist in powers of mutual impenetrability or resistance, might have helped facilitate Newton's realization that mass can be considered at a point. Or at least, because they lack parts outside parts, such bodies would not stand in the way of that realization.

References

Bennett, J. and Remnant P. (1978). "How Matter Might First be Made". New Essays on Rationalism and Empiricism. Ed. C. E. Jarrett, J. King- Farlow, and F. J. Pelletier. *Canadian Journal of Philosophy*, Supplementary Volume 4: 1-11.

Cohen, I.B. 'Versions of Isaac Newton's first published paper', *Archives Internationales d'Histoire des Sciences*, 11, 1958: 357-75.

---

has abstracted away only something that is necessary to body, not everything, since bodies also have the power to produce sensations. He is saying that condition (3) is necessary to body-hood, but he is not saying that it is sufficient.

[43] *De gravitatione*, 27; Cf. *Principia*, 942.

[44] Gorham, January 2011, 24.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

Cohen, I. B. (1999). A guide to Newton's Principia (with contributions by M. Nauenberg, & G. Smith). In I. Newton, *The Principia: Mathematical principles of natural philosophy* (I. B. Cohen, & A. Whitman, Trans.) (pp. 1–370). Berkeley: University of California Press.

Conn, Christopher H. "Two Arguments for Lockean Four-Dimensionalism". *British Journal for the History of Philosophy*, ISSN 0960-8788, 10/1999, Volume 7, Issue 3, pp. 429 – 446.

Dempsey, Liam, "Written in the flesh: Isaac Newton on the mind–body relation", *Studies in History and Philosophy of Science*, Vol. 37, No.3 (2006), pp. 420-441.

Descartes, R. (1985). *The philosophical writings of Descartes*, Vols. 1 & 2 (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.). Cambridge: Cambridge University Press. (Abbreviated CSM).

Gabbey, Alan. (2011) 'The term *materia* in Newton and in the Newtonian tradition', in *Materia: XIII Colloquio Internazionale, Roma, 7-8-9 gennaio 2010*. Atti a cura di Delfina Giovannozzi e Marco Veneziani, pp. 423-445. Firenze: Leo S. Olschki Editore. Lessico Intellettuale Europeo 113.

Garber, Daniel (2001). *Descartes Embodied*, Cambridge: Cambridge University Press.

Gorham, G. (2011a) "How Newton Solved the Mind-Body Problem". *History of Philosophy Quarterly* 28: 21-44.

Gorham, G. (2011b) "Newton on God's Relation to Space and Time: The Cartesian Framework". *Archiv für Geschichte der Philosophie*, 93: 281-320.

Grant, Edward. *Much Ado about Nothing: Theories of Space and Vacuum from the Middle Ages to the Scientific Revolution*. New York: Cambridge University Press, 1981.

Henry, John: "A Cambridge Platonist's Materialism: Henry More and the Concept of Soul", *Journal of the Warburg and Courtauld Institutes*, Vol. 49 (1986), pp. 172-195.

Hill, B. (2003). Newton's De Gravitatione et Aequipondio Fluidorum and Lockean Four-Dimensionalism". *British Journal for the History of Philosophy* 11: 309-321.

Hill, J., 2004, "Locke's Account of Cohesion and its Philosophical Significance," *British Journal for the History of Philosophy*, 12 (4): 611 – 630.

Hooke, Robert: Lectures *Potentiae Restitutiva*, or of Spring, Explaining the Power of Springing Bodies. Printed for John Martyn, Printer to the Royal Society, at the Bell in St. *Pauls* Church-Yard, 1678.

Kochiras, H. (2009). Gravity and Newton's Substance Counting Problem, *Studies in History and Philosophy of Science*, 40(3): 267-280.

Presented June 14, 2012 at the 7th Quadrennial Fellows Conference of the Pittsburgh Center for Philosophy of Science, convened at Mugla University

Kochiras, H. (2011). "Gravity's Cause and Substance Counting: Contextualizing the Problems", *Studies in History and Philosophy of Science,* 42(1): 167-184

Kochiras, H. "Spiritual Presence and Dimensional Space beyond the Cosmos", *Intellectual History Review*, 22(1) 2012: 41–68.

Locke, John, *An Essay Concerning Human Understanding*, edited Peter H. Nidditch, New York: Oxford
University Press, 1975.

McGuire, J. E. (1978). "Existence, Actuality and Necessity: Newton on Space and time". *Annals of Science*, 35, 463-508.

McGuire, J.E. and Slowik, Edward, "Newton's Ontology of Omnipresence and Infinite Space", *Oxford Studies in Early Modern Philosophy* (forthcoming).

More, Henry. *Philosophical Writings of Henry More*, ed. F.I. MacKinnon. New York: Oxford University Press, 1925.

Newton, I. (2004). Newton: Philosophical writings (A. Janiak, Ed.). Cambridge: Cambridge University Press.

Newton, I. (1962). *Unpublished scientific papers of Isaac Newton* (A. R. Hall, & M. B. Hall, Eds. & Trans.). Cambridge: Cambridge University Press.

Newton, I. (1959–1971). *Correspondence of Isaac Newton* (H. W. Turnbull, J. F. Scott, A. R. Hall, & L. Tilling, Eds.) (7 vols.). Cambridge: Cambridge University Press.

Newton, I. (1952). *Opticks, or A treatise of the reflections, refractions, inflections and colors of light.* Based on the fourth edition of 1730. New York: Dover.

Nolan, Lawrence, "Reductionism and Nominalism in Descartes's Theory of Attributes". *Topoi* 16: 129–140, 1997.

Reid, Jasper, "The Spatial Presence of Spirits among the Cartesians" *Journal of the History of Philosophy*, Volume 46, Number 1, January 2008, pp. 91-117.

Slowik, Edward. Newton's Metaphysics of Space: A "Tertium Quid" betwixt Substantivalism and Relationism, or Merely a "God of the (Rational Mechanical) Gaps"? *Perspectives on Science*, Volume 17, Number 4, Winter 2009, pp. 429-456.

Stein, Howard, "Newton's Metaphysics", in I. Bernard Cohen and George Smith, editors, *The Cambridge Companion to Newton*, Cambridge: Cambridge University Press, 2002.

Arto Siitonen

On Reichenbach's Dissertation from 1916

Preface

The German philosopher Hans Reichenbach (1891 - 1953) is well known as one of the representatives of the stream of thought called 'logical empiricism'. It is less known that in his youth he was a devoted Kantian philosopher. This can be seen in his doctoral dissertation that he defended at the university of Erlangen on March 2, 1915. The title of this work was *Der Begriff der Wahrscheinlichkeit für die mathematische Darstellung der Wirklichkeit*, and it was published 1916 in the journal *Zeitschrift für Philosophie und philosophische Kritik*.

In his *curriculum vitae* (p. 2 of the dissertation), Reichenbach tells that he has studied philosophy, mathematics, physics and pedagogy in Berlin, Munich, and Göttingen. He mentions among his teachers Ernst Cassirer, Max Planck, Alois Riehl, Carl Stumpf, Ernst von Aster, David Hilbert and Edmund Husserl. He characterizes himself as "a Kantian philosopher".

As far as I know, Reichenbach's dissertation has not been translated into any other language. A direct translation of its title in English would be: *The Concept of Probability for the Mathematical Presentation of Reality*. In what follows below, is an analysis and commentary of the main lines of thought of

Reichenbach's dissertation.

1. Subjectivism vs. objectivism

Reichenbach vigorously argues in favour of the *objective* interpretation of the concept of probability. Probability belongs to reality and not only to our knowledge of reality. He considers the question of probability part of the debate on the basic concepts of our knowledge of nature. The *subjectivists* have given up their belief in objective knowledge and consider science a game of human thoughts, whereas the *objectivists* rely on the real validity of scientific results. The question for objectivists concerns the issue: which of the necessary elements of knowledge are characterized as true? Kant's critique of reason is a method for the systematic study of this question. However, in spite of his great discovery, until recently only a few critical studies have been dedicated to the concepts employed in the positive sciences. Only a few philosophers have followed up their inquiries in mathematics and physics, and used their critical acumen with the methods constantly applied therein.

As Reichenbach sees it, the analysis of the concept of probability is divided between the claims of the exact science of probability calculus and those of the unclear and vaguely applied concepts employed in everyday life. Philosophical investigation has too often started from the latter, thereby admitting neither a precise statement of the problem nor its solution. Moreover, this form of inquiry has led to subjectivism.

.

His dissertation has the following structure: The *first chapter* "Das Problem" presents the basic difficulty and its treatment in recent philosophy. The *second chapter* gives an analysis of specific problems of probability. It is divided into four sections that concern the so-called "probability machine", games of chance, the theorem of combined probabilities, and the theory of errors. The *third chapter* is entitled "Deduktion des Wahrscheinlichkeitsprinzips" (Deduction of the Principle of Probability), and the *fourth chapter* studies the relation of probability judgments to reality.

Reichenbach thought that there was a curious contrast between *probability* and the principle of *causality*, to the effect that when the causality principle is inapplicable, the connections between the phenomena are accommodated under the concept of probability. This leads to a *probability paradox*: it seems that only because we are unable to determine the specific connections must we be content to suppose that the case under investigation is probable or improbable. However, may such a supposition, conditioned only by our ignorance, be expressed as a claim of objective validity?

Reichenbach goes on to consider the subjective interpretation of probability as represented by Carl Stumpf, and the objective interpretation as represented by Johannes von Kries. According to Stumpf, probability concerns the relation of the positive to the possible. A probability judgment presents the given state of our knowledge. Stumpf's concept cannot give a measure of reasonable expectation, because such a

measure must say something about real things and not only about knowledge. Reichenbach remarks ironically that it may be that our expectation is regulated in accordance with our knowledge, whereas real things are certainly not thus regulated (cf. p. 6 of his dissertation).

von Kries formulated the following four principles: (1) probability judgments are true or false, (2) they give a definite structure to reality, (3) they contain a prediction of future occurrences, and (4) what is expressed in them is based on a non-empirical principle, that of variability within a given scope ('*Spielräume*'). Reichenbach's main criticism of von Kries' analysis is that it yields only a subjective certainty to the principle of causality. Von Kries does not think that there is an objective foundation for predictions.

> The correctness of applying the principle of causality has been proven by KANT in his *Critique*; if this had not been possible, then we would not be justified to draw a subjective certainty from this principle. (p. 11).[1]

2. Probability judgments

The philosophers F. A. Lange, E. F. Apelt, A. Fick and Kurt Grelling represent the following view: probability judgments are disjunctive or hypothetical, and the *implicans* of the hypothetical judgment gives the total area of a condition whereas the *implicate* expresses a certain conclusion. Probability is then identifiable with the relation between the extensions of the expressions contained in the judgment. Reichenbach says of Fick's view of probability judgments:

> Fick is right in calling these sentences, like mathematical sentences, synthetic judgments a priori. GRELLING has followed him in this conception.
>
> These mathematical sentences are not confirmable by experience; but whether they are applicable to reality is a special problem, which has to be solved outside the discipline of probability calculus through philosophical inquiry. (p. 12).[2]

Fick formulated the problem as the assumption that there seems to be an asymptotic relation between probability and reality. Reichenbach stresses the important role that probability plays in everyday life, in games of chance, in insurance firms and in modern physics. All this makes it reasonable to suppose that probability laws are *objective laws of nature*, their validity being philosophically justifiable. For a general strategy of research, he suggests the following: let us act *as if* the validity of probability laws were already proven, and let us then study, on the basis of this supposition, what kind of organization has thus been presupposed. This is the natural route to a critique of reason, he claims, because the mathematician presupposes as valid the principles that the philosopher seeks to put under criticism.

3. Special problems of probability

After having stated the *general problem* to be dealt with, Reichenbach turns to an analysis of *special problems* of probability, taking a "probability machine", games of chance, the theorem of combined probabilities, and the theory of errors of measurement as examples.

The probability machine is an idealization (comparable to the more famous machine that Alan Turing invented later). It features a rolling band in which a piston moving in a cylinder strikes holes. The band consists of a sheet with regularly changing white and black stripes. The probability distribution for white and black holes is 1:2. But why is this? The conditions are: (i) that the piston strikes a hole very frequently, (ii) that both of the occurrences (the rolling of the band and the striking the hole) are mutually independent, (iii) that these occurrences have a joint effect, and (iv) that there exists a *probability function* in accordance with which the occurrences can be classified.

The probability function can be given a form which contains the number of the values x for the time of striking (between an arbitrary interval from a to b), and N for the total number of the times of striking. The conditions under which probability sentences can be applied, are: (1) it must be decidable which regularity underlies the occurrences in the event that the results of the probability calculation are correct, (2) the regularity must be observable through its physical effects. While the second condition is empirically ascertainable, the existence of a probability function is not an empirical affair; rather, the question here is of "a metaphysical principle of the knowledge of nature" ("ein metaphysisches Prinzip der Naturerkenntnis", p. 26).

As in the case of the probability machine, one has to suppose the existence of a probability function when confronted with games of chance (Reichenbach refers to roulette), with combined probabilities, and with

measurement errors. The necessity for this implicit presupposition leads to a philosophical question: how do I know that there is such a function? It is the task of experience to determine the special form of such a function, but no experience can teach *that* a probability function exists for any thinkable case. Observation can only tell us something of a finite number of cases.

The situation is similar to the case of *causality*: its special content is always empirically given, but no observation tells us that what is observed in a single case is universally valid, or that from now on any case whatsoever will comply with this law. However, the regularity in the case of *probability* is different from causal regularity in that it concerns occurrences that are *not* causally combined. The *law of probability* states the following: in cases in which the law of causality is inapplicable, the principle of probability is valid.

4. Justification of the law of probability

One may wonder how, if the law of probability does not stem from experience, it is to be *justified*. The fourth chapter, "Deduction of the principle of probability", is devoted to this question. The basic premise is the same as in Kant's *transcendental deduction* of the fundamental principles of knowledge – among them the principle of causality, challenged by Hume. According to Reichenbach, Kant was able to prove that the axioms of geometry are *synthetic judgments a priori*. Because geometry was an

established branch of science, the *a priori* foundation Kant gave to it was considered as valid as geometry itself. However, the situation is different with the theory of errors and the theory of games of chance – which do not have the same reputation as geometry. It thus became Reichenbach's task to find corresponding validation of the probability principle.

Reichenbach's strategy was basically the same as Kant's in his transcendental deduction. Reichenbach sought to show that the necessity of probability function follows from the idea of knowledge of nature; that probability theory is justified due to its connection to other principles of knowledge and to the unity of knowledge. He even called his procedure "transcendental proof" (p. 48).

The judgments of mathematics are hypothetical, whereas those of physics (or, in general, reality judgments) are categorical. The former are given us through pure intuition, the latter are determined by empirical intuition and given through perception. In spite of their empirical character, reality judgments contain an uneliminable *apriori core,* due to which perceptions are embedded in a net of relations. The terms of these relations are not empirical concepts, but contain syntheses of further relations. In other words, the Kantian 'transcendental unity of apperception' is not limited to judgments, but also applies to their elements. Synthesis through transcendental apperception is "the highest point" in Kant's analysis of knowledge. Reichenbach points out to page A 109/B 134 of *Critique of Pure Reason* in this connection, and concerning the constitution of an empirical object to page A 116/B 146.

The principle of transcendental synthesis is that an ideal structure is thought to be applicable to real occurrences. Accordingly, judgments of physics maintain that certain mathematical structures are true of given segments of reality. Nevertheless, one is not allowed to say that the mathematical structure and reality correspond to each other, because real occurrences are always determined by infinitely many such structures. What happens is that certain objects of empirical intuition are *coordinated* with the equations of mathematics: for instance, certain kinds of gases are thus coordinated with Boyle's law $p \cdot V = R \cdot T$. It is through the mathematical formulas that the quantitative dimension of the objects is determined. Without this supposition, empirical knowledge in general would be impossible.

5. Mathematics and physics

Mathematical equations are ideal structures; they signify the relations between the objects of pure intuition. Physical judgments apply mathematical equations and use them to give an approximative presentation of reality. They do not exhaust reality – empirical objects also contain uneliminatory irrational elements. The theories of modern physics are systems of equations. The possibility of physical knowledge means the possibility of giving numerical approximations. Physics proceeds in two directions: towards concrete, singular phenomena by way of continuous *specialization*, and towards more and more *general* laws and theories by

transforming the former constants of nature into functions.

That mathematics can be applied to real objects is not an empirical sentence, but rather the methodical presupposition of physics – this means nothing else than the great basic idea in Kant's theory of knowledge. *The principle of probability* plays an uneliminable role in the application of mathematical equations to reality. It contains the idea that there is, for a series of repetitions of the same magnitude, a probability function. The probability principle is "an objective law of nature" (p.71) that is necessarily valid for the occurrences of nature. For instance, when we throw a standard six-sided dice, the probability principle states that there is a finite number N of throws to the effect that the distribution 1:6 for each of the sides is reached within limits that are not larger than an arbitrary number e for deviation from the distribution. In accordance with *Bernouilli's theorem*, each of the numbers 1-6 appears as frequently as any other.

6. Probability and reality

The last chapter (Chapter 4 of the dissertation) concerns the relation of probability judgments to reality. At the outset, Reichenbach expresses his indebtness to Kant as follows:

> Thus the existence of the probability function has been deduced in the sense in which KANT uses the word deduction in his transcendental philosophy. The necessity of such a law is in the last instance only to be grasped by an insight, and it has therefore been called a synthetic judgment a priori; it is not logically derivable from other principles of knowledge. But in this deduction has been

> shown, how this law is connected to the whole knowledge of nature in general; and while it has been shown that that principle means a necessary condition of all physical knowledge, its validity concerning experience has been *proven*. (p. 65).[3]

Reichenbach claims that the existence of the probability function cannot be deduced from other principles of knowledge. The judgement that suggests that there exists a probability function is *synthetic a priori*.

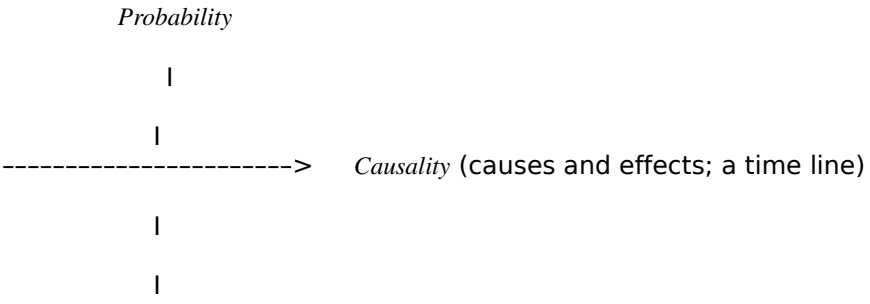
> It is even the criterion of *a priori* laws that they cannot be confirmed or refuted by any special experience but are rather forms of ordering that make experience possible. Experience in the scientific sense is a presentation of reality connecting the given contents of perception in the sense of stable, *a priori* ordering forms. (p. 70).[4]

No experience can contain anything to contradict the principle of probability, just as deviation from the law of causality can never appear. We can neither prove nor disprove these laws empirically. An experiment designed in order to confirm the principle of causality would presuppose the very principle, just as an experiment to prove that the probability function exists already presupposes that it does. If the law of the distribution of approximate values were to contain a definite numerical value for 'N', it would lose its *a priori* character and become a special law of nature, confirmable or refutable by experiment. On the other hand, we interpret apparent deviations from the law as a conformity to the law. Although there may be, say, thirty throws of 6 in a row, we would say that in the long run the probability of getting 6 is 1:6. In the case of a continuous anomaly, we would rather suspect that the dice is unbalanced than call the probability principle into question.

Accordingly, the principle of probable distribution of values is an objective

law of nature, which is valid with certainty. That there exist probabilities is itself not another probability, then, claims Reichenbach, and some law must be valid with certainty in order for some other to be valid with probability. It is probable that single instances conform to the law of probability distribution; it is certain that with a growing number of cases, an approximation to this law will be reached with evolving accuracy.

The relation between the principles of *causality* and *probability* is a complementary one: when the probability principle is added to the causality principle, knowledge of nature becomes possible. The former connects phenomena together in a *vertical* direction, the latter *horizontally*, as it were (cf. pp. 62 and 73 of the dissertation). Let us display this by the following figure:



Reichenbach considered that he had thus clarified the relation of probability judgments to reality. Fick has shown that the sentences of probability calculus are *synthetic* sentences *a priori*, presenting a system that is analogous to the sentences of geometry. According to Reichenbach, it has been shown that these sentences must necessarily be true of reality, i.e. that real things are necessarily subordinated to them. This

subordination is due to the principle of the probability function, which is an objective law of nature. It is because of this proof that the *parallelism between the principles of probability and of geometry* has become complete.

At the end of his dissertation, Reichenbach remarks that theoretical physics in its modern form is essentially based on probability considerations, it being a further task to explicate the underlying philosophical principles behind the law of entropy and behind the Maxwell-Bolzmannian statistics.

7. Comments

The following features of Reichenbach's dissertation deserve to be underlined:

(1) One cannot overestimate the thoroughly *Kantian spirit* of the book. It accepts the idea of *synthetic a priori* principles of knowledge. Its basic principle, that of the probability function, is presented as independent of experience and as necessarily valid for all future experience. Moreover, it offers an *aprioristic perspective* on all knowledge. Although reality judgments are distinguished from judgments of mathematics, they are considered *a priori* judgments in so far as their structural core is *a priori*. Their empirical part is their contingent content.

Reichenbach sees knowledge as a synthetic product of our conceptual and

perceptual capacities, as Kant does. He also uses Kant's *transcendental method* of reasoning. He accepts Kant's view of geometry and applies his leading question to probability: *how is the probability function possible?* He even claims, in the spirit of Kant's transcendental dialectics, that reality is partly irrational: our knowledge can only approximate but never exhaust it.

What is most remarkable in respect of Kantianism is that Reichenbach in fact *complements* Kant's analysis of categories with the probability function. Causality is one of the categories in Kant's transcendental analytic; in stressing that probability is independent of causality, Reichenbach makes it an extra category. One could accordingly see Reichenbach's early philosophy as *strengthened Kantianism*. His subsequent progress towards pure empiricism required a profound change and transformation of the ideas of his dissertation period. Perhaps one could call the philosophy presented in the dissertation *Kantianism with a probability accentuation*.

(2) Reichenbach's dissertation presents a strong case in favour of the objectivistic interpretation of probability, according to which probability is an objective feature of reality, independent of our suppositions and of our knowledge. Knowledge is based on probability, rather than the other way around. This idea remained Reichenbach's conviction throughout his career. He gave it more and more precise articulation in order to attack all forms of subjectivism. He always thought, as he argues in the introduction to his dissertation, that if the subjective interpretation were allowed, the idea of scientific knowledge would be given up. It is possible that this view

became one of the sources of his conflict with Kantianism, because Kantianism stresses the importance of the world of appearances, a world constituted by human perceptual and rational capacities.

(3) The laws of causality and probability are harmoniously connected in the dissertation. Both are *a priori* and neither is superordinated to the other. In this respect, there was a change in Reichenbach's later philosophy when he took the concept of causality under closer scrutiny. In doing this, he continued Hume's dissolution of the causality principle – an undertaking that took him further away from Kantian ideas.

(4) Reichenbach's stresses in his criticism of Stumpf that the probability calculus has to give us a measure of reasonable expectation ("ein Mass der vernünftigen Erwartung", p. 6) in respect to future. This would yield a foundation for predictions. Likewise, in his criticism of Kries he put it in question, whether Kries' probability judgment does express anything of the future reality ("zukünftigen Wirklichkeit", p. 9). He was to return to these themes – notably in his work *Experience and Prediction*. He was also later to seize the explication task of the law of entropy and the Maxwell-Boltzmann statistics, indicated at the end of his dissertation. This he did in his last, unfinished, work *The Direction of Time*.

Translated Citations as these Appear in the Original Text

1) "Die Rechtmässigkeit der Anwendung des Kausalprinzips ist von KANT in der transzendentalen Deduktion der Kritik dargetan worden; wäre dies nicht möglich gewesen, so hätten wir nicht das Recht, aus diesem Prinzip

eine subjektive Gewissheit zu entnehmen." ((1916), p. 11).

2) "FICK hat Recht, wenn er alle diese Sätze wie die mathematischen Sätze synthetische Urteile a priori nennt. In dieser Auffassung ist ihm GRELLING gefolgt.

Diese mathematischen Sätze sind durch die Erfahrung nicht zu bestätigen; ob sie aber auf die Wirklichkeit anwendbar sind, ist ein besonderes Problem, das jenseits der Disziplin der Wahrscheinlichkeitsrechnung durch philosophische Untersuchung gelöst werden muss." (p. 12).

3) "Es ist somit die Existenz einer Wahrscheinlichkeitsfunktion deduziert worden in dem Sinne, wie KANT das Wort Deduktion für die Transzendentalphilosophie gebraucht. Die Notwendigkeit einer solchen Gesetzmässigkeit lässt sich letzten Endes nur einsehen, und sie ist deshalb ein synthetisches Urteil a priori genannt worden; sie lässt sich nicht logisch aus anderen Grundsätzen der Erkenntnis ableiten. In dieser Deduktion aber ist gezeigt worden, wie jenes Gesetz im Zusammenhang steht mit der gesamten Naturerkenntnis überhaupt; und indem dargetan worden ist, dass jenes Prinzip eine notwendige Bedingung aller physikalischen Erkenntnis bedeutet, ist seine Gültigkeit von der Erfahrung *bewiesen* worden." (p. 65).

4) "Das ist gerade das Kriterium apriorischer Gesetze, dass sie nicht durch irgendeine spezielle Erfahrung bestätigt oder widerlegt werden können, sondern die vorher gesetzten Formen der Einordnung bilden, die erst die spezielle Erfahrung möglich machen. Erfahrung im wissenschaftlichen Sinne ist eine solche Darstellung der Wirklichkeit, die die gegebenen Wahrnehmungsinhalte im Sinne fester apriorischer Ordnungsformen zusammenfügt." (p. 70).

Choice of Literature (cf. pp. 78 f)

E. F. Apelt: *Theorie der Induktion*. Leipzig 1854. W. Engelmann.

Ernst Cassirer: *Substanzbegriff und Funktionsbegriff*. Berlin 1910. B. Cassirer.

A. Fick: *Philosophischer Versuch über die Wahrscheinlichkeiten*. Würzburg 1883.

Kurt Grelling: *Die philosophischen Grundlagen der Wahrscheinlichkeitsrechnung*. Abhandlungen der Friesschen Schule III. Band. 3. Heft. 1910.

Immanuel Kant: *Kritik der reinen Vernunft*. 2. Aufl. Akademieausgabe.

Johannes von Kries: *Die Prinzipien der Wahrscheinlichkeitsrechnung*. Freiburg 1886. J.C.B. Mohr.

F.A. Lange: *Logische Studien*. Leipzig 1894. Baedeker.

Carl Stumpf: *Über den Begriff der mathematischen Wahrscheinlichkeit*. Sitzungsbericht der philosophisch-historischen Klasse der königlich bayerischen Akademie der Wissenschaften zu München. 1892.
– *Über die Anwendung des mathematischen Wahrscheinlichkeitsbegriffes auf Teile eines Continuums*. Ebenda.