Explanatory Power and Explanatory Justice

Jonah N. Schupbach and Jan Sprenger

February 14, 2014

Abstract

Crupi and Tentori (2012) propose a condition of adequacy for any Bayesian measure of explanatory power, which they call Explanatory Justice. They criticize a measure recently defended by Schupbach and Sprenger(2011) for failing to satisfy this condition, and they offer a new *explanatorily just* measure of explanatory power. In this paper, we investigate Explanatory Justice's merits as a condition of adequacy. We offer three arguments against this condition, thus supporitng the idea that a measure of explanatory power should rather be *unjust*. Then, in order to ensure that the debate advances beyond a mere battle of intuitions, we offer some new results derived from the empirical study described in (Schupbach, 2011a). All of this strengthens the case for Schupbach and Sprenger's measure while simultaneously posing new challenges to Crupi and Tentori's alternative proposal.

In "The Logic of Explanatory Power" (Schupbach and Sprenger, 2011), we defend a probabilistic explication of explanatory power by means of two related lines of argument. First, we show that certain conditions of adequacy are uniquely satisfied by a particular class of ordinally equivalent explicata. Second, we prove several theorems for this class of functions, showing that such measures nicely fit certain explanatory intuitions. Recently, both lines of argument have come under fire. Crupi and Tentori (2012) call into question all but one of the conditions of adequacy that drive our uniqueness theorem. Additionally, they argue vigorously against one of the theorems that follows from our explication, and in favor of a condition of adequacy for such explications that runs contrary to this theorem. Crupi and Tentori reject our account and replace it with their own proposed explication. After briefly reviewing the relevant results from our original paper (section 1), we respond to each of their criticisms in turn (section 2). Along the way, we argue against Crupi and Tentori's alternative proposal.

1 The Logic of Explanatory Power, Revisited

Though we are not explicit about this, the work that we undertake in our (2011) is most properly characterized as a straightforward attempt at Carnapian explication. Following the description of this method in (Carnap, 1950, ch. 1), we begin by formulating our problem with an "informal explanation" – intended, in Carnap's words, "to make clear what is meant as the explicandum." Inspired especially by Peirce, but also noticing a common theme running through several contemporary accounts of the nature of explanation, we clarify our explicandum as follows (p. 108):

> The sense of explanatory power that this article seeks to analyze [read explicate] has to do with a hypothesis's ability to decrease the degree to which we find the explanandum surprising (i.e., its ability to increase the degree to which we expect the explanandum). More specifically, a hypothesis offers a powerful explanation of a proposition, in this sense, to the extent that it makes that proposition less surprising.

We then pursue a probabilistic explication of this particular concept of explanatory power. In doing so, we seek an explicatum that satisfies,

to some satisfactory extent, each of Carnap's four desiderata: simplicity, precision, similarity, and fruitfulness.[1] In order to ensure that it will be both functionally *simple* and mathematically *precise*, we introduce a specific formal framework for the desired explicatum. This results in our first condition of adequacy (where E generically denotes any probabilistic measure of explanatory power):[2]

CA1: There exists an analytic function g such that $E_{Pr}(e,h) = g[Pr(h|e), Pr(h|\neg e), Pr(e)]$. Values of $E_{Pr}(e,h)$ range in $[-1,+1]$.

Next, in order to ensure some degree of *similarity* between our intuitive notion of explanatory power and the desired explicatum, we lay down three more substantive conditions of adequacy.[3] Each of these is intended to hold our explicatum accountable to intuitions pertaining to the target notion of explanatory power, so that results derived from our explicatum truly do have some bearing on that concept:

CA2: Ceteris paribus, the greater the degree of statistical relevance between e and h, the greater the value of $E_{Pr}(e,h)$.

CA3: If h_2 is probabilistically independent from e, h_1, and their conjunction (i.e., $Pr(e \wedge h_2) = Pr(e)Pr(h_2)$, $Pr(h_1 \wedge h_2) = Pr(h_1)Pr(h_2)$, and $Pr(e \wedge h_1 \wedge h_2) = Pr(e \wedge h_1)Pr(h_2)$), then $E_{Pr}(e,h_1) = E_{Pr}(e,h_1 \wedge h_2)$.

CA4: If $\neg h$ entails e, then the values of $E_{Pr}(e,h)$ do not depend on the values of $Pr(h)$. Formally, there exists a function f so that, if $\neg h \vDash e$, then either $E_{Pr}(e,h) = f[Pr(h|e)]$ or $E_{Pr}(e,h) = f[Pr(e)]$.

With these conditions of adequacy in place, we prove our central Theorem 1: All measures satisfying CA1-4 are monotonically increasing functions of the "posterior ratio", $Pr(h|e)/Pr(h|\neg e)$. A straightforward implication of this theorem is that any function that satisfies CA1-4 will be some

[1] There has been a recent surge of interest in Carnap's notion of explication. The following works all include more extensive descriptions of Carnap's method than we have space to give here: (Boniolo, 2003; Eagle, 2004; Maher, 2007; Kitcher, 2008; Schupbach, 2011b; Justus, 2012; Shepherd and Justus, 2012).

[2] Throughout the paper, in accordance with our original article, we assume that the propositions e, h, and so on are contingent, and we require that Pr be regular. For the sake of continuity with the paper that we are responding to, we adopt Crupi and Tentori's presentation of our conditions.

[3] We leave a defense of our explicatum's fruitfulness as something to be pursued in future work; see (Schupbach, 2011b, ch. 5) for such a defense.

member of a class of functions, all of which are ordinally equivalent to $Pr(h|e)/Pr(h|\neg e)$. Our Theorem 2 then selects one particular explicatum out of this class, making use of a set of additional conditions of adequacy. The result is our adoption of the following measure of explanatory power (specifically denoted by $\mathcal{E}$):

$$\mathcal{E}_{Pr}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)}.$$

As a further line of argument in defense of our explication, we also prove four theorems, each purportedly showing that our explicatum fits well with our intuitive judgments pertaining to the intended notion of explanatory power. Crupi and Tentori only take issue with the first of these (Theorem 3 in the original paper), which can be stated as follows:

Theorem 3. If $Pr(e^*|e \wedge h) = Pr(e^*|e)$ and $Pr(e^*|e) \neq 1$, then:

- if $\mathcal{E}_{Pr}(e,h) > 0$, then $\mathcal{E}_{Pr}(e,h) > \mathcal{E}_{Pr}(e \wedge e^*,h) > 0$,
- if $\mathcal{E}_{Pr}(e,h) < 0$, then $\mathcal{E}_{Pr}(e,h) < \mathcal{E}_{Pr}(e \wedge e^*,h) < 0$, and
- if $\mathcal{E}_{Pr}(e,h) = 0$, then $\mathcal{E}_{Pr}(e,h) = \mathcal{E}_{Pr}(e \wedge e^*,h) = 0$.

2 On Crupi and Tentori's Objections

Crupi and Tentori's objections can be organized into two categories: those that challenge the conditions of adequacy underlying our first uniqueness theorem (Theorem 1), and an objection to the intuitive merits of Theorem 3. After briefly responding to challenges in the first category, we focus more extensively on Crupi and Tentori's argument against Theorem 3.

2.1 Objections to our Conditions of Adequacy

Against CA1, Crupi and Tentori write:

> [I]t is not the case that all posterior ratio measures satisfy CA1-CA4, and this is especially due to CA1, which goes beyond constraining the ordinal structure. [...] By this restrictive character, CA1 prevents Schupbach and Sprenger's main result from being a proper representation theorem for posterior ratio measures.

Crupi and Tentori's main point just seems to be that a proper representation theorem for posterior ratio measures would necessarily give a representation of *all* such measures. Since our uniqueness theorem does not give an axiomatization for *all* posterior ratio measures, it does not properly represent all such measures. Moreover, Crupi and Tentori apparently chalk this up as a serious defect, so much so that they take it upon themselves to find alternative axiomatic foundations for the complete class of posterior ratio measures.

If this is a fair characterization of Crupi and Tentori's first criticism, then it seems to be based upon a misunderstanding of our project. The goal of our Theorem 1 is not to find a general representation theorem for posterior ratio measures. Rather, the goal, as clarified in the original paper (e.g., see p. 107), is to introduce a set of conditions that we desire to be true of our explicatum, and then to show that these desiderata only hold for a particular family of probabilistic measures. The unique family of satisfactory measures turns out to be a proper subset of the family of posterior ratio measures. Crupi and Tentori point to the fact that our theorem singles out a subclass of posterior ratio measures as a defect. However, they give us no argument for this conclusion; i.e., they give us no reason to agree with them that a successful theorem from desirable constraints to a unique family of posterior ratio measures must be a proper representation theorem for *all* posterior ratio measures.

It is worth clarifying that our uniqueness theorem *is* a proper representation theorem, not for all posterior ratio measures, but for all posterior ratio measures that are analytic and range in $[-1, +1]$. Crupi and Tentori may or may not take issue with these additional stipulations; if they do, they have not made that clear nor again have they given us the relevant arguments. Regardless, the important issue is not whether our mathematical restrictions might rule out plausible candidate explicata. Rather, the important question is whether one can find an explicatum that has a specified, desirable structure, which satisfies Carnap's desiderata. Accordingly, our past work provides us with a ready-made response to those who would argue that our restrictions on the formal structure of E are too strong, insofar as this work supports the thesis that there exists an explicatum with this strong mathematical structure that satisfies Carnap's desiderata.

As already mentioned, this complaint motivates Crupi and Tentori to develop a general representation theorem for posterior ratio measures,

and this is potentially very helpful. As they write (p. 369), "A clear and general axiomatization of posterior ratio measures fosters insight into their distinctive properties and thus a focused discussion of their implications." Specifically then, it is useful to have such a theorem in hand especially because an alternative formal representation of posterior ratio measures reveals certain characteristics and implications of our explicatum. And these potentially afford us with a further check on the intuitive merits of our explication.

Nonetheless, we are doubtful that Crupi and Tentori's representation theorem can provide us with such aid. The reason is found in the following condition of adequacy, which drives their theorem:[4]

Posteriors. There exists a function g such that $E_{Pr}(e,h) = g[Pr(h|e), Pr(h|\neg e)]$.

This condition is simply too coarse to give us any new intuitions regarding the plausibility of our preferred class of measures. Posteriors is just obviously true of our explicatum and its intuitive virtues or counterintuitive vices are already made apparent by the very statement of $\mathcal{E}$.

Regarding CA2-4, we can be very brief. Crupi and Tentori take no issue with CA3. Against CA2, they write, "CA2 is left rather unspecified, as statistical relevance can be measured in various ways. This is innocuous for their proof. Yet, to allow for a discussion of the content and plausibility of this assumption, a sharper rendition would be helpful." But what is to be gained by such a move? As Crupi and Tentori acknowledge, our proof of Theorem 1 goes through, regardless of how one chooses to measure statistical relevance. It seems to us that there is real value in highlighting this fact by refusing to describe CA2 in terms of a particular measure.

Finally, Crupi and Tentori are absolutely right to point to CA4 as the weak spot in our account. We point this out ourselves in the original article. The hope is that future work interacting with our article inspects more deeply the merits of this condition of adequacy. Unfortunately, Crupi and Tentori merely point to our earlier observation that this condition

[4]We should note that Crupi and Tentori never claim that Posteriors is an illuminating constraint on an explication of explanatory power. In fact, they take this condition to be a precisification of an earlier statement of our own CA4 (found in a previous draft of our paper). It is true that this is one way of precisifying that earlier condition. The precisification that we favor, however, is the statement of CA4 that made its way into the final, printed version of our article.

lacks intuitive motivation (additionally, there do not seem to be any clear intuitions that weigh against this condition).

2.2 On Explanatory Justice

The above criticisms of our conditions of adequacy are relatively minor. The real heart of Crupi and Tentori's disagreement with our work centers around our Theorem 3. We have stated this theorem formally above; in less formal terms, it describes the general effect that adding irrelevancies to the explanandum has on the explanatory power that a hypothesis has over that explanandum. The idea is simply motivated and stated in the following passage (Schupbach and Sprenger, 2011, p. 115):

> [A]s the evidence becomes less statistically relevant to some explanatory hypothesis h (with the addition of irrelevant propositions), it ought to be the case that the explanatory power of h relative to that evidence approaches the value at which it is judged to be explanatorily irrelevant to the evidence ($\mathcal{E} = 0$). Thus, if $\mathcal{E}_{Pr}(e,h) > 0$ [in which case e and h are positively relevant to one another, or $Pr(e|h) > Pr(e)$], then this value should decrease with the addition of e^* [a proposition that is irrelevant to h in the light of e] to our evidence: $0 < \mathcal{E}_{Pr}(e \wedge e^*, h) < \mathcal{E}_{Pr}(e,h)$. Similarly, if $\mathcal{E}_{Pr}(e,h) < 0$ [in which case, $Pr(e|h) < Pr(e)$], then this value should increase with the addition of e^*: $0 > \mathcal{E}_{Pr}(e \wedge e^*, h) > \mathcal{E}_{Pr}(e,h)$. And finally, if $\mathcal{E}_{Pr}(e,h) = 0$ [in which case, $Pr(e|h) = Pr(e)$], then this value should remain constant at $\mathcal{E}_{Pr}(e \wedge e^*, h) = 0$.

Theorem 3 thus clarifies that, by adding irrelevant propositions to our explanandum, we *dilute* explanatory power (be it positive or negative); the more irrelevant to h our explanandum becomes through this process, the closer it approaches that value at which we interpret h as explanatorily irrelevant to the explanandum ($\mathcal{E} = 0$).

Contrary to this, Crupi and Tentori argue that *positive* degrees of explanatory power should indeed be diluted in these cases, but not so with *negative* degrees of explanatory power. They formalize this intuition in their "Explanatory Justice" condition:

Explanatory Justice. If e^* is probabilistically independent from e, h, and their conjunction, then:

i) if $E_{Pr}(e,h) > 0$, then $E_{Pr}(e \wedge e^*,h) < E_{Pr}(e,h)$, and

ii) if $E_{Pr}(e,h) \leq 0$, then $E_{Pr}(e \wedge e^*,h) = E_{Pr}(e,h)$.

The remainder of this section compares the relative merits of our Theorem 3 and Explanatory Justice.

Recall that, in the formulation of our problem, following Carnap's example, we give an "informal explanation" of our target concept of explanatory power. We clarify there that the sense of "explanatory power" that we explicate has to do with "a hypothesis's ability to decrease the degree to which we find the explanandum surprising (i.e., its ability to increase the degree to which we expect the explanandum)." Additionally, we show that this sense of the term is commonly applied in scientific reasoning, contemporary philosophy, as well as in everyday life (Schupbach and Sprenger 2011, p. 108; see also, Schupbach 2011b, sections 1.1 and 2.3).

Crupi and Tentori recognize that this is our target concept of interest. Moreover, they accept the same informal explanation for the sense of "explanatory power" that they aim to examine. So, for example, regarding their proposed alternative measure of explanatory power (specifically denoted $\mathcal{E}^*$), they write (p. 375), "$\mathcal{E}^*(e,h)$ conveys in probabilistic terms a view of explanatory power that is conceptually sound, as it transparently involves how the background surprisingness / expectedness of explanandum e is reduced by assuming candidate explanans h." Given that both sides of this debate agree on this much, the central question to investigate when adjudicating between Theorem 3 and Explanatory Justice is which of these, if either, fits better with the concept of explanatory power as increase in expectedness.

Two simple observations favor Theorem 3. First, the effect that some h has on the degree to which we expect a proposition will be more or less significant depending on how relevant h is to that proposition. The more positively [negatively] relevant h is to the proposition, the more accepting h would lead us to expect the proposition to be true [false]; only when h is entirely irrelevant to the proposition in question will accepting h not affect the degree to which we expect that proposition to be true or false. Second, the conjunction that results from tacking propositions that are irrelevant to h (in the light of e) on to some explanandum e that *is* relevant to h is, on the

whole, less relevant to h than e taken by itself.[5] In other words, statistical relevance (positive or negative) between propositions is gradually diluted as irrelevancies are conjoined to one of those propositions. These two observations together imply the following conclusion: the more irrelevant propositions one loads into an explanandum, the less effect h will have on the expectedness of the conglomerate. Given that our goal is to explicate the notion of explanatory power as increase in expectedness / decrease in surprise, this conclusion just amounts to Theorem 3. In order to defend Explanatory Justice, Crupi and Tentori would somehow have to show that the above argument holds for cases of positive relevance, but not for cases of negative relevance.

The above argument provides a clear and compelling case for requiring that positive and negative degrees of explanatory power ought to be similarly diluted when irrelevancies are tacked onto the explanandum. Why then do Crupi and Tentori think that negative explanatory power ought *not* be so diluted? They give the following reason (p. 370): "[S]hould it be the case that $[0 >] E_{Pr}(e \wedge e^*, h) > E_{Pr}(e, h)$ in these circumstances, then one would be allowed to indefinitely relieve a lack of explanatory power, no matter how large, by adding more and more irrelevant explananda, simply at will." But this remark is misleading in at least two different ways. First, the negative explanatory power that h has over e is not at all mitigated by considerations of irrelevancies; $E_{Pr}(e, h)$ is what it is, and the value of this is, of course, not affected by considerations of e^*. It is only once we change the explanandum (e.g., to $e \wedge e^*$) that we see explanatory power shift. Thus, referring to a case presented by Crupi and Tentori, one cannot make the hypothesis that <the coin is fair> a good explanation, or even a less bad explanation of 10 tails being consecutively flipped by considering a host of irrelevant statements.

Second, it is misleading to say that one could indefinitely relieve a lack of explanatory power for any measure that satisfies Theorem 3. "Lack of explanatory power" is ambiguous. On the one hand, one might interpret this phrase to mean *negative* degree of explanatory power. On this reading, to relieve a lack of explanatory power is to increase negative degrees of explanatory power toward the neutral value of 0. On the other hand, this phrase may be interpreted such that any hypothesis with a positive,

[5] Proposition e^* is irrelevant to h in the light of e if and only if $Pr(e^*|h \wedge e) = Pr(e^*|e)$; see also our formulation of Theorem 3.

but weak degree of explanatory power over the explanandum substantially lacks explanatory power. In this case, an "indefinite relief of a lack of explanatory power" means unrestricted increases in explanatory power toward the maximal value of 1. It would indeed be embarrassing if our explication allowed that one could relieve explanatory power in this second sense by tacking irrelevancies onto the explanandum; but, thankfully, this is not the case. Quite to the contrary, on our account, *positively* explanatory hypotheses are rendered *worse* by adding such irrelevancies.

Theorem 3 clarifies that any relief that comes by tacking irrelevancies onto the explanandum is of the first type. As we argue above, this result is actually desirable. For further motivation, consider a simple coin scenario. Let h be <the coin is fair>, and let the initial e be <the first 10 flips of this coin yield 10 consecutive tails>. Now imagine that we conjoin irrelevant propositions to the explanandum until we have, say, $1,000$ statements about rolls of various fair *dice* tacked on to e. Should not the explanatory power of h over this massive conglomerate be rather close to the interpretive point at which we say that h is explanatorily irrelevant to our explanandum? Our explication, but not Crupi and Tentori's, has this property. Of course, the explanatory power that h has over this changing explanandum never actually reaches the point of explanatory irrelevance, and it never becomes positive.[6]

There is another good reason to require that explanatory power be symmetrically diluted. Negative degree of explanatory power between h and e is interpreted as positive degree of explanatory power between h and $\neg e$. Thus, for example, the interpretation corresponding to a minimal degree of explanatory power (e.g., $\mathcal{E}_{Pr}(e,h) = -1$) is that h maximally explains e's *falsity* (i.e., it maximally explains $\neg e$). More generally, according to our explication, h fails to explain e to the extent that it explains $\neg e$. This is formalized in our "Symmetry" condition of adequacy, which states (2011, p. 113):

Symmetry. $E_{Pr}(e,h) = -E_{Pr}(\neg e,h)$.

[6]This is in accordance with similar results in the related fields of confirmation and coherence theory. In these fields, the statistical relevance of a proposition h with regard to a conjunction of proposition is usually measured as an *aggregate* (e.g., Fitelson, 2003; Meijs, 2005): that is, the overall value lies in between the most extreme values, and it is not determined by the most extreme value of relevance between h and a single conjunct alone, as Crupi and Tentori suggest.

Crupi and Tentori interpret negative explanatory power in the same way, as is manifest by their favorably citing our motivation for Symmetry ("the less surprising / more expected the truth of e is in light of a hypothesis, the more surprising / less expected is e's falsity"), and then by their introducing their own (weaker) symmetry condition (p. 372). But if one accepts this as the appropriate interpretation of negative degrees of explanatory power (and it is difficult to think of another way in which such degrees could be interpreted), then one will want negative explanatory power to be diluted in just the same way as positive explanatory power.

Why? Stipulate that h has negative explanatory power over e – i.e., it has positive explanatory power over $\neg e$. Now conjoin to e an irrelevant proposition e^*. If one requires, as Crupi and Tentori do, that $E(e \wedge e^*, h) = E(e, h)$, then we may also conclude that $E(\neg(e \wedge e^*), h) = E(\neg e, h)$ (this follows from Symmetry but also from Crupi and Tentori's weakened version of that constraint). This amounts to saying that h explains the disjunction $\neg e \vee \neg e^*$ to the same degree that it explains $\neg e$ by itself. (Or, put directly in terms of expectedness, h boosts our expectedness for $\neg e \vee \neg e^*$ to the same extent that it boosts our expectedness for $\neg e$ taken by itself.[7]) For example, if h states that a particular coin is strongly biased towards heads, e describes the outcome "tails" of a toss of that coin, and e^* describes the result of rolling an unbiased *die* (let us assume it comes up with a six), then the relevant conditions of probabilistic independence are satisfied. In that case, Crupi and Tentori's account implies that h explains $\neg e$ (the hypothetical outcome "heads") to the same degree that it explains $\neg e \vee \neg e^*$ (the hypothetical outcome "heads or the die comes up with a value between 1 and 5"). That is, by accepting Explanatory Justice, Crupi and Tentori also accept that irrelevant *disjunctions* do not dilute positive degrees of explanatory power.[8] But this is arguably going too far since the disjuncts may not stand in any probabilistic relevance relation to the explanans.

So we conclude that, quite to the contrary of Crupi and Tentori's recommendation, any explicatum that plausibly corresponds to the target

[7] It is important to keep the distinction between *expectedness* and *increase in expectedness* in mind when considering this result. Of course, $\neg e \vee \neg e^*$ should always be more expected than $\neg e$ taken by itself – whether or not one accepts h. This is simply because the former is logically weaker than the latter. But our point here is that, when $\neg e$ is positively relevant to h and $\neg e^*$ is irrelevant to h, h will affect a larger *increase* in the expectedness of $\neg e$ than the *increase* affected in the expectedness of $\neg e \vee \neg e^*$.

[8] If e^* is probabilistically independent of e, h, and their conjunction, then $\neg e^*$ is also independent of $\neg e, h$, and their conjunction.

explicandum of explanatory power ought to break with Explanatory Justice. Note that it would be one thing if Crupi and Tentori had another sense of "explanatory power" in mind than the one that we describe by appealing to the notion of shifts in expectedness (or surprise). However, they are clear that this is the explicandum that they intend to study. Given this, what we have tried to argue here is that, keeping this target notion strictly in mind leads one to affirm that explanatory power ought *not* be thought of as explanatorily just in Crupi and Tentori's sense.

Finally, before moving on to the next section of this paper, we need briefly to face one more criticism. In a footnote, Crupi and Tentori (2012, pp. 372-373, fn. 10) state the idea behind our Theorem 3 in terms of the following condition:[9]

(W): If e^* is probabilistically independent from e, h, and their conjunction, then $|E_{Pr}(e \wedge e^*, h)| < |E_{Pr}(e, h)|$.

Crupi and Tentori then point out that our explication does not imply (W) because "it gives $\mathcal{E}(e \wedge e^*, h) = \mathcal{E}(e, h) = -1$ whenever $h \vDash \neg e$." On this basis, they conclude that "posterior ratio measures do not convey the idea of [dilution] in a fully coherent fashion."

However, the idea of dilution should not hold without restriction. It must be balanced against another intuitive idea having to do with the conditions under which our explicatum should take its minimum value. We state this idea as follows (2011, p. 111):

Minimality. $E_{Pr}(e, h)$ takes minimal value if and only if $h \vDash \neg e$.

With our specific explanandum firmly in mind, we argue that this requirement is sensible as follows: "h should be minimally explanatory of e if e is maximally surprising in the light of h, and this occurs whenever h implies the falsity of e." Of course, if h implies the falsity of e, then it also implies the falsity of $e \wedge e^*$; thus, Minimality clearly compels us to require that negative explanatory power not be diluted in this extreme case.

To strike a balance between the idea of dilution and Minimality, then, what we really want is a clever formal means to rule out cases in which

[9]Note that (W) does not adequately capture the sense of dilution. This is primarily because it could be the case that $|E_{Pr}(e \wedge e^*, h)| < |E_{Pr}(e, h)|$ and simultaneously $E_{Pr}(e \wedge e^*, h)$ may have a different sign than $E_{Pr}(e, h)$. Note also the discrepancy between the independence conditions set out in the antecedent of (W) as compared to those set out in the antecedent of our Theorem 3. This discrepancy will prove important to the present discussion.

$\mathcal{E}_{Pr}(e,h) = -1$ from the statement of Theorem 3. Here, it is important to highlight a discrepancy between the relations of independence required in the antecedent conditions of (W) and those required in the antecedent conditions of our Theorem 3. Crupi and Tentori require that "e^* be probabilistically independent from e, h, and their conjunction" whereas we require that "$Pr(e^*|e \wedge h) = Pr(e^*|e)$". It turns out that our independence condition, but not Crupi and Tentori's, effectively rules out those cases in which $\mathcal{E}_{Pr}(e,h) = -1$. This is simply because, as clarified by Minimality, $\mathcal{E}_{Pr}(e,h) = -1$ if and only if $h \models \neg e$. In such cases, assuming the standard treatment of conditional probabilities, the term $Pr(e^*|e \wedge h)$ is undefined. Accordingly, there is no sense in which we can ask whether $Pr(e^*|e \wedge h) = Pr(e^*|e)$, and so Theorem 3 simply does not speak to such cases. Thus, while our explicatum may not imply Crupi and Tentori's explication of dilution (W), it *does* imply our Theorem 3 (as we proved in Appendix C of our paper).

2.3 Comparing $\mathcal{E}$ and $\mathcal{E}^*$

As part of their response to our paper, Crupi and Tentori put forward an alternative explication of explanatory power. They show that the following measure – along with any other measure ordinally equivalent to this one – will satisfy Explanatory Justice:

$$\mathcal{E}^*(e,h) = \begin{cases} \dfrac{Pr(e|h) - Pr(e)}{1 - Pr(e)} & \text{if } Pr(e|h) \geq Pr(e) \\ \dfrac{Pr(e|h) - Pr(e)}{Pr(e)} & \text{if } Pr(e|h) < Pr(e) \end{cases}$$

In the foregoing section, we have argued that intuitions about our target explicandum of explanatory power favor Theorem 3 over Explanatory Justice as a desirable constraint on an adequate explicatum. Accordingly, we have already put forward a case against explicating explanatory power via $\mathcal{E}^*$. However, we readily admit that the relevant intuitions are not easy to come by, and we are certainly open to the possibility that others have intuitions that conflict with our own. In this section then, we would like to advance the debate beyond the purely theoretical level by relating some recent empirical findings to our disagreement with Crupi and Tentori.

(Schupbach, 2011a) presents an empirical study comparing the descriptive merits of several candidate explications of explanatory power.[10] In this study, experimental participants are asked to provide a series of judgments of explanatory power in a ball-and-urn, chance context. Each such judgment relates how well participants think one of two hypotheses (h_A or h_B) explains a body of evidence (e); here, we denote any such judgment $J(e, h_{A/B})$. Schupbach then compares these judgments to corresponding theoretical results derived from each particular candidate measure of explanatory power. These results are derived both using subjective probabilities collected from the participants and using objective probabilities calculated from the chance setup.

In order to test which explication is most similar to human judgments of explanatory power, Schupbach analyzes the experimental results in two ways: (1) For each measure, the Euclidean distance between participant judgments and derived, theoretical results is calculated; (2) Each measure's residual distribution (i.e., the distribution of values of $J(e, h) - E(e, h)$ corresponding to any measure E) is examined and compared. On the basis of this study, Schupbach argues that the measure of explanatory power $\mathcal{E}$ provides the best fit of all candidate measures in terms of both analyses. I.e., theoretical results derived from $\mathcal{E}$ sit closer, on average, to participant judgments of explanatory power, and the mean residual corresponding to $\mathcal{E}$ is closer to the ideal value of zero than that corresponding to any other measure (indeed, this is the only mean residual that does not differ significantly from zero).

Importantly, given that this study was run prior to Crupi and Tentori's recent contribution to the field, Schupbach's considered list of candidate measures of explanatory power does not include measure $\mathcal{E}^*$. In the remainder of this section then, we present the results of including this measure in the previous study. It turns out that both types of analysis – (1) and (2) above – strongly favors $\mathcal{E}$ over Crupi and Tentori's $\mathcal{E}^*$.

First, $\mathcal{E}$ scores better than $\mathcal{E}^*$ in terms of simple Euclidean distance from participant judgments (results of this comparison are shown in Table 1). Here, numbers represent the average distance that theoretical results are from participant judgments – hence, of course, the smaller the number the better. In fact, Popper's (1959) measure $E_P(e, h) = [Pr(e|h) -$

[10]The design of these experiments is based closely on a chance setup previously applied by Phillips and Edwards (1966) and more recently by Crupi and Tentori themselves (Tentori et al., 2007) in their comparison of various Bayesian measures of *confirmation*.

Measure	Distance from $J(e, h_A)$	Distance from $J(e, h_B)$
Subjective probabilities:		
$\mathcal{E}$	5.597	5.211
$\mathcal{E}^*$	7.272	7.577
Objective probabilities:		
$\mathcal{E}$	5.617	6.218
$\mathcal{E}^*$	7.104	7.183

Table 1. Distances between participant judgments and measures.

$Pr(e)]/[Pr(e|h) + Pr(e)]$ as well as the various finitely rescaled versions of Good (1960) and McGrew's (2003) measure $E_M(e, h) = ln[Pr(e|h)/Pr(e)]$ considered in Schupbach's original empirical study also score better than $\mathcal{E}^*$ in this regard.

Second, we can also compare the distribution of residuals calculated with respect to each measure. Note that, for an explication that is able to fit participant judgments very well without systematically underestimating or overestimating such judgments, we expect the distribution of such residuals to have a mean value of zero. Again, $\mathcal{E}$ comes out as better than $\mathcal{E}^*$ in this regard (results of this comparison are shown in Table 2). In fact, every other candidate probabilistic explication of explanatory power that Schupbach considers in the original study does significantly better than $\mathcal{E}^*$.[11] The next worst is the difference measure $E_D(e, h) = Pr(e|h) - Pr(e)$ with mean residuals of $-.098$ using subjective probabilities and $-.095$ using objective probabilities. What we learn from this is that, on average, the results derived using $\mathcal{E}^*$ greatly underestimate people's judgments of explanatory power. Contrasted with this is $\mathcal{E}$, which is the only candidate measure that provides results that do not, on average, differ significantly from the ideal value of zero (Schupbach, 2011a, p. 825).

3 Conclusions

Crupi and Tentori initiate a dialogue on probabilistic explications of explanatory power that we wholeheartedly welcome. Their article proposes an interesting, new requirement for such explications, while simultaneously offering several challenging criticisms of our recent explication. In

[11] $p < .00001$ in a paired t-test comparing the distributions of residuals.

Measure	Mean Residual	σ
Subjective probabilities:		
$\mathcal{E}$	-.015	.335
$\mathcal{E}^*$	.237	.395
Objective probabilities:		
$\mathcal{E}$	.071	.361
$\mathcal{E}^*$	.208	.391

Table 2. Sample statistics.

this paper, after briefly responding to some of their more minor objections (in section 2.1), we turned to the disagreement over their Explanatory Justice requirement. We suggested that the key question to ask, with regards to this disagreement, is whether or not explanatory power should be diluted toward the point of explanatory irrelevance when irrelevancies are added to the explanandum. Crupi and Tentori argue that negative explanatory power should not be diluted in such circumstances, even if positive explanatory power should be. In this paper, we contend that explanatory power should be symmetrically diluted in such circumstances – regardless of whether it is positive or negative.

To this end, in section 2.2, we argued that if one keeps the specific target explicandum of explanatory power in mind, intuitions favor a symmetrical dilution requirement (Theorem 3). If Crupi and Tentori remain committed to Explanatory Justice, they may well want to consider whether they have a (perhaps subtly) different explicandum in mind – their assertions to the contrary notwithstanding.

Next, in section 2.3, we went further by showing how past empirical work bears on this debate. We compared the descriptive merits of Crupi and Tentori's favored explanatorily just explicatum $\mathcal{E}^*$ with our own measure $\mathcal{E}$ and came to two conclusions: (1) $\mathcal{E}$ comes closer than $\mathcal{E}^*$, on average, to experimental participants' judgments pertaining to explanatory power; and (2) $\mathcal{E}^*$'s corresponding mean residual differs from the ideal value of zero to a greater extent than any other considered, candidate measure of explanatory power (whereas $\mathcal{E}$'s corresponding mean residual comes the closest to zero). This evidence suggests that the explication that Crupi and Tentori propose, unlike our own, is actually quite dissimilar to people's working concept of explanatory power.

References

Boniolo, G. (2003). Kant's explication and Carnap's explication: The redde rationem. *International Philosophical Quarterly*, 43(3):289–298.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Crupi, V. and Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science*, 79(3):365–385.

Eagle, A. (2004). Twenty-one arguments against propensity analyses of probability. *Erkenntnis*, 60:371–416.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, 63(3):194–199.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(2):319–331.

Justus, J. (2012). Carnap on concept determination: Methodology for philosophy of science. *European Journal for Philosophy of Science*, 2(2):161–179.

Kitcher, P. (2008). Carnap and the caterpillar. *Philosophical Topics*, 36(1):111–127.

Maher, P. (2007). Explication defended. *Studia Logica*, 86:331–341.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science*, 54:553–567.

Meijs, W. (2005). *Probabilistic Measures of Coherence*. PhD thesis, Erasmus Universiteit Rotterdam.

Phillips, L. D. and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3):346–354.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London.

Schupbach, J. N. (2011a). Comparing probabilistic measures of explanatory power. *Philosophy of Science*, 78(5):813–829.

Schupbach, J. N. (2011b). *Studies in the Logic of Explanatory Power*. PhD thesis, University of Pittsburgh, Pittsburgh.

Schupbach, J. N. and Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, 78(1):105–127.

Shepherd, J. and Justus, J. (2012). X-phi, explication, and formal epistemology. Unpublished Manuscript.

Tentori, K., Crupi, V., Bonini, N., and Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, 103:107–119.