The Logic of Explanatory Power

Jonah N. Schupbach and Jan Sprenger*†

[Forthcoming in *Philosophy of Science*]

Abstract

This paper introduces and defends a probabilistic measure of the explanatory power that a particular explanans has over its explanandum. To this end, we propose several intuitive, formal conditions of adequacy for an account of explanatory power. Then, we show that these conditions are uniquely satisfied by one particular probabilistic function. We proceed to strengthen the case for this measure of explanatory power by proving several theorems, all of which show that this measure neatly corresponds to our explanatory intuitions. Finally, we briefly describe some promising future projects inspired by our account.

*To contact the authors, please write to: Jonah N. Schupbach, Department of History and Philosophy of Science, University of Pittsburgh, 1017 Cathedral of Learning, Pittsburgh, PA 15260; email: jonah@schupbach.org.

†We would like to thank David Atkinson, John Earman, Theo Kuipers, Edouard Machery, John Norton, Jeanne Peijnenburg, Jan-Willem Romeijn, Tomoji Shogenji, and especially Stephan Hartmann for helpful comments on earlier versions of this article. Jan Sprenger would like to thank audiences at PROGIC 09, Groningen, and the ESF Workshop in Woudschouten who all gave him valuable feedback on presentations of this work. Jonah N. Schupbach would like to thank Tilburg University's Center for Logic and Philosophy of Science (TiLPS) which supported him with a research fellowship during the time that he worked on this article.

1 Explanation and Explanatory Power

Since the publication of Hempel and Oppenheim's (1948) classic investigation into "the logic of explanation," philosophers of science have earnestly been seeking an analysis of the nature of explanation. Necessity (Glymour, 1980), statistical relevance (Salmon, 1971), inference and reason (Hempel and Oppenheim, 1948; Hempel, 1965), familiarity (Friedman, 1974), unification (Friedman, 1974; Kitcher, 1989), causation (Woodward, 2003), and mechanism (Machamer et al., 2000) are only some of the most popular concepts that such philosophers draw upon in the attempt to describe necessary and sufficient conditions under which a theory explains some proposition.[1] A related project that is, on the other hand, much less often pursued by philosophers today is the attempt to analyze the *strength* of an explanation – i.e., the degree of *explanatory power* that a particular explanans has over its explanandum. Such an analysis would clarify the conditions under which hypotheses are judged to provide *strong versus weak* explanations of some proposition, and it would also clarify the meaning of comparative explanatory judgments such as "hypothesis *A* provides a better explanation of this fact than does hypothesis *B*."

Given the nature of these two projects, the fact that the first receives so much more philosophical attention than the second can hardly be explained by appeal to any substantial difference in their relative philosophical imports. Certainly, the first project has great philosophical significance; after all, humans on the individual and social levels are constantly seeking and formulating explanations. Given the ubiquity of explanation in human cognition and action, it is both surprising that this concept turns out to be so analytically impenetrable,[2] and critical that philosophers continue to strive for an understanding of this notion. The second project is, however, also immensely philosophically important. Humans regularly make judgments of explanatory power and then use these judgments to develop preferences for hypotheses, or even to infer outright to the truth of certain hypotheses. Much of human reasoning – again, on individual and social levels – makes use of judgments of explanatory power. Ultimately then, in order to understand and evaluate human reasoning generally, philosophers need to come to a better understanding of explanatory power.

The relative imbalance in the amount of philosophical attention that these two projects receive is more likely due to the prima facie plausible but ultimately unfounded assumption that one must have an analysis of explanation

[1] See Woodward (2009) for a recent survey of this literature.

[2] Lipton (2004, 23) refers to this fact that humans can be so good at doing explanation while simultaneously being so bad at describing what it is they are doing as the "gap between doing and describing."

before seeking an analysis of explanatory power. This assumption is made compelling by the fact that in order to analyze the strength of something, one must have some clarity about what that thing is. However, it is shown to be much less tenable in light of the fact that humans *do* generally have some fairly clear intuitions concerning explanation. The fact that there is no consensus among philosophers today over the precise, necessary and sufficient conditions for explanation does not imply that humans do not generally have a firm semantic grasp of the concept of explanation. Just how firm a semantic grasp of this concept humans actually have is an interesting question. One claim of this paper will be that our grasp of the notion of *explanation* is at least sufficiently strong to ground a precise formal analysis of *explanatory power* – even if it is not strong enough to determine a general account of the nature of explanation.

This paper attempts to bring more attention to the second project above by formulating a Bayesian analysis of *explanatory power*. Moreover, the account given here does this without committing to any particular theory of the nature of explanation. Instead of assuming the correctness of a theory of explanation and then attempting to build a measure of explanatory power derivatively from this theory, we begin by laying out several, more primitive adequacy conditions that, we argue, an analysis of explanatory power should satisfy. We then show that these intuitive adequacy conditions are sufficient to define for us a unique probabilistic analysis and measure of explanatory power.

Before proceeding, two more important clarifications are necessary. First, we take no position on whether our analysis captures the notion of explanatory power *generally*; it is consistent with our account that there be other concepts that go by this name, but which do not fit our measure.[3] What we *do* claim, however, is that our account captures at least one familiar and epistemically compelling sense of *explanatory power* that is common to human reasoning.

Second, because our explicandum is the strength or power *of an explanation*, we restrict ourselves in presenting our conditions of adequacy to speaking of theories that do in fact provide explanations of the explanandum in question.[4] This account thus is not intended to reveal the conditions under which

[3] As a possible example, Salmon (1971) and Jeffrey (1969) both argue that there is a sense in which a hypothesis may be said to have positive *explanatory power* over some explanandum so long as that hypothesis and explanandum are statistically relevant, regardless of whether they are negatively or positively statistically relevant. As will become clear in this paper, insofar as there truly is such a notion of explanatory power, it must be distinct from the one that we have in mind.

[4] To be more precise, the theory only needs to provide a *potential* explanation of the explanandum – where a theory offers a *potential explanation* of some explanandum just in case, if it were true, then it would be an *actual explanation* of that explanandum. In other words, this account may be used to measure the strength of any potential explanation, regardless of whether the

a theory is explanatory of some proposition (that is, after all, the aim of an account of explanation rather than an account of explanatory power); rather, its goal is to reveal, for any theory *already known to provide such an explanation*, just how strong that explanation is. Ultimately then, this paper offers a probabilistic logic of explanation that tells us the explanatory power of a theory (explanans) relative to some proposition (explanandum), *given that that theory constitutes an explanation of that proposition*. In this way, this paper forestalls the objection that two statements may stand in the probabilistic relation described while not simultaneously constituting an explanation.

2 The Measure of Explanatory Power $\mathscr{E}$

The sense of *explanatory power* that this paper seeks to analyze 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. This sense of explanatory power dominates statistical reasoning where scientists are "explaining away" surprise in the data by means of assuming a specific statistical model; e.g., in the omnipresent linear regression procedures. But the explaining hypotheses need not be probabilistic; for example, a geologist will accept a prehistoric earthquake as explanatory of certain observed deformations in layers of bedrock to the extent that deformations of that particular character, in that particular layer of bedrock, *et cetera* would be less surprising given the occurrence of such an earthquake.

This notion finds precedence in many classic discussions of explanation. Perhaps its clearest historical expression occurs when Peirce (1935, 5.189) identifies the explanatoriness of a hypothesis with its ability to render an otherwise "surprising fact" as "a matter of course."[5] This sense of explanatory power

explanans involved is actually true.

[5]This quote might suggest that explanation is tied essentially to *necessity* for Peirce. However, elsewhere, Peirce clarifies and weakens this criterion: "to explain a fact is to show that it is a necessary *or, at least, a probable result* from another fact, known or supposed" (Peirce, 1935, 6.606, emphasis mine). See also (Peirce, 1958, 7.220).

There are two senses in which our notion of explanatory power is more general than Peirce's notion of explanatoriness: first, a hypothesis may provide a powerful explanation of a surprising proposition, in our sense, and still not render it a matter of course; i.e., a hypothesis may make a proposition much less surprising while still not making it unsurprising. Second, our sense of explanatory power does not suggest that a proposition must be surprising in order to be explained; a hypothesis may make a proposition much less surprising (or more expected) even if the latter is not very surprising to begin with.

may also be seen as underlying many of the most popular accounts of explanation. Most obviously, DN and IS accounts (Hempel, 1965), and necessity accounts (Glymour, 1980) explicitly analyze explanation in such a way that a theory that is judged to be explanatory of some explanandum will necessarily increase the degree to which we expect that explanandum.

Our formal analysis of this concept proceeds in two stages: In the first stage, a parsimonious set of adequacy conditions is used to determine a measure of explanatory power up to ordinal equivalence. In other words, we show that, for all pairs of functions f and f' that satisfy these adequacy conditions, $f(e,h) > (=,<) f(e',h')$ if and only if $f'(e,h) > (=,<) f'(e',h')$; all such measures thus impose the same *ordinal* relations on judgments of explanatory power. This is already a substantial achievement. In the second stage, we introduce more adequacy conditions in order to determine a unique measure of explanatory power (from among the class of ordinally equivalent measures).

In the remainder, we make the assumption that the probability distribution is regular; i.e., only tautologies and contradictions are awarded rational degrees of belief of 1 and 0. This is not strictly required to derive the results below, but it makes the calculations and motivations much more elegant.

2.1 Uniqueness Up to Ordinal Equivalence

The first adequacy condition is, we suggest, rather uncontentious. It is a purely formal condition intended to specify the probabilistic nature and limits of our analysis:

CA1 (Formal Structure**):** For any probability space and regular probability measure $(\Omega, \mathscr{A}, Pr(\cdot))$, $\mathscr{E}$ is a measurable function from two propositions $e,h \in \mathscr{A}$ to a real number $\mathscr{E}(e,h) \in [-1,1]$. This function is defined on all pairs of contingent propositions; i.e., cases such as $Pr(e) = 0$ etc. are not in the domain of $\mathscr{E}$.[6] This implies by Bayes's Theorem that we can represent $\mathscr{E}$ as a function of $Pr(e)$, $Pr(h|e)$ and $Pr(h|\neg e)$, and we demand that any such function be analytic.[7]

The next adequacy condition specifies, in probabilistic terms, the general notion of explanatory power that we are interested in analyzing. As mentioned,

[6] The background knowledge term k always belongs to the right of the solidus "|" in Bayesian formalizations. Nonetheless, here and in the remainder of this paper, we choose for the sake of transparency and simplicity in exposition to leave k implicit in all formalizations.

[7] A real-valued function f is analytic if we can represent it as the Taylor expansion around a point in its domain. This requirement ensures that our measure will not be composed in an arbitrary or ad-hoc way.

a hypothesis offers a powerful explanation of a proposition, in the sense that we have in mind, to the extent that it makes that proposition less surprising. In order to state this probabilistically, the key interpretive move is to formalize a decrease in surprise (or increase in expectedness) as an increase in probability. This move may seem dubious depending upon one's interpretation of probability. Given a physical interpretation (e.g., a relative frequency or propensity interpretation), it would be difficult indeed to saddle such a psychological concept as surprise with a probabilistic account. However, when probabilities are themselves given a more psychological interpretation (whether in terms of simple degrees of belief or the more normative degrees of *rational* belief), this move makes sense. In this case, probabilities map neatly onto degrees of expectedness.[8] Accordingly, insofar as surprise is inversely related to expectedness (the more surprising a proposition, the less one expects it to be true), it is straightforwardly related to probabilities. Thus, if h decreases the degree to which e is surprising, we represent this with the inequality $Pr(e) < Pr(e|h)$. The strength of this inequality corresponds to the degree of statistical relevance between e and h giving us:

CA2 (Positive Relevance**):** Ceteris paribus, the greater the degree of statistical relevance between e and h, the higher $\mathscr{E}(e,h)$.

The third condition of adequacy defines a point at which explanatory power is unaffected. If h_2 does nothing to increase or decrease the degree to which e, h_1, or any logical combination of e and h_1 are surprising, then $h_1 \wedge h_2$ will not make e any more or less surprising than h_1 by itself already does. In this case, tacking h_2 on to our hypothesis has no effect upon the degree to which that hypothesis alleviates our surprise over e. Given that explanatory power has to do with a hypothesis's ability to render its explanandum less surprising, we can state this in other words: if h_2 has no *explanatory power* whatever relative to e, h_1, or any logical combination of e and h_1, then explanandum e will be explained no better nor worse by conjoining h_2 to our explanans h_1. Making use of the above probabilistic interpretation of a decrease in surprise, this can be stated more formally as follows:

CA3 (Irrelevant Conjunction**):** If $Pr(e \wedge h_2) = Pr(e) \times Pr(h_2)$ and $Pr(h_1 \wedge h_2) = Pr(h_1) \times Pr(h_2)$ and $Pr(e \wedge h_1 \wedge h_2) = Pr(e \wedge h_1) \times Pr(h_2)$, then $\mathscr{E}(e, h_1 \wedge h_2) = \mathscr{E}(e, h_1)$.

[8] This is true by definition for the first, personalist interpretation; in terms of the more normative interpretation, probabilities still map neatly onto degrees of expectedness, though these are more specifically interpreted as *rational* degrees of expectedness.

The following adequacy condition observes that explanatory power, in our sense, does not depend upon the prior plausibility of the explanans. This is because the extent to which an explanatory hypothesis alleviates the surprising nature of some explanandum does not depend on considerations of how likely that hypothesis is in and of itself. Rather, to decide the effect of a hypothesis upon the surprise (expectedness) of some explanandum, one compares how surprising (expected) the explanandum is apart from considerations of the hypothesis to how surprising (expected) it would be *granting the truth of the hypothesis*. In making this specific comparison, it is simply not necessary (and not helpful) to know how plausible the explanatory hypothesis is on its own. With this sense of explanatory power in mind then, it is perfectly sensible to talk about two hypotheses that are vastly unequal in their respective plausibilities having the same amount of explanatory power over an explanandum. For example, dehydration and cyanide poisoning may be (approximately) equally powerful explanations of symptoms of dizziness and confusion insofar as they both make such symptoms less surprising to the (approximately) same degree. And this is true despite the fact that dehydration is typically by far the more plausible explanans. In light of these considerations, we require the following

CA4 (Irrelevance of Priors**):** Values of $\mathscr{E}(e,h)$ do not depend upon the values of $Pr(h)$.[9]

These four conditions allow us to derive the following theorem (proof in **Appendix 1**):

Theorem 1. *All measures of explanatory power satisfying* **CA1-CA4** *are monotonically increasing functions of the posterior ratio* $Pr(h|e)/Pr(h|\neg e)$.

From this theorem, two important corollaries follow. First, we can derive a result specifying the conditions under which $\mathscr{E}$ takes its maximal and minimal values (proof in **Appendix 1**):

Corollary 1. $\mathscr{E}(e,h)$ *takes maximal value if and only if h entails e, and minimal value if and only if h implies* $\neg e$.

Note that this result fits well with the notion of explanatory power that we are analyzing, according to which a hypothesis explains some proposition to the

[9]The following weaker version of **CA4** actually suffices in the proof of **Theorem 1**: When either h or $\neg h$ implies e, values of $\mathscr{E}(e,h)$ and $\mathscr{E}(e,\neg h)$ do not depend upon the values of $Pr(h)$ and $Pr(\neg h)$. Nonetheless, the notion of explanatory power that we analyze motivates the condition that explanatory power does not depend upon $Pr(h)$ generally – not merely when h or $\neg h$ implies e. Accordingly, we include this stronger condition here.

extent that it renders that proposition less surprising (more expected). Given this, any h ought to be *maximally* explanatorily powerful regarding some e when it renders e maximally unsurprising (expected), and this occurs whenever h guarantees the truth of e ($Pr(e|h) = 1$). Similarly, 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 ($Pr(e|h) = 0$).

The second corollary constitutes our desired ordinal equivalence result:

Corollary 2. *All measures of explanatory power satisfying* **CA1-CA4** *are ordinally equivalent.*

To see why the corollary follows from the theorem, let r be the posterior ratio of the pair (e,h), and let r' be the posterior ratio of the pair (e',h'). Without loss of generality, assume $r > r'$. Then, for any functions f and f' that satisfy **CA1-CA4**, we obtain the following inequalities:

$$f(e,h) = g(r) > g(r') = f(e',h') \qquad f'(e,h) = g'(r) > g'(r') = f'(e',h'),$$

where the inequalities are immediate consequences of **Theorem 1**. So any f and f' satisfying **CA1-CA4** always impose the same ordinal judgments, completing the first stage of our analysis.

2.2 Uniqueness of $\mathscr{E}$

This section pursues the second task of choosing a specific and suitably normalized measure of explanatory power out of the class of ordinally equivalent measures determined by **CA1-CA4**. To begin, we introduce an additional, purely formal requirement of our measure:

CA5 (Normality and Form**):** $\mathscr{E}$ is the ratio of two functions of $Pr(e \wedge h)$, $Pr(\neg e \wedge h)$, $Pr(e \wedge \neg h)$ and $Pr(\neg e \wedge \neg h)$, each of which are homogeneous in their arguments to the least possible degree $k \geq 1$.[10]

Representing $\mathscr{E}$ as the ratio of two functions serves the purpose of normalization. $Pr(e \wedge h)$, $Pr(\neg e \wedge h)$, $Pr(e \wedge \neg h)$ and $Pr(\neg e \wedge \neg h)$ fully determine the probability distribution over the truth-functional compounds of e and h, so it is appropriate to represent $\mathscr{E}$ as a function of them. Additionally, the requirement that our two functions be "homogenous in their arguments to the least possible degree $k \geq 1$" reflects a minimal and well-defined simplicity assumption akin to those advocated by Carnap (1950) and Kemeny and Oppenheim

[10] A function is *homogeneous* in its arguments to degree k if its arguments all have the same total degree k.

(1952, 315). This assumption effectively limits our search for a unique measure of explanatory power to those that are the most cognitively accessible and applicable.

Of course, larger values of $\mathscr{E}$ indicate greater explanatory power of h with respect to e. $\mathscr{E}(e,h) = 1$ (being $\mathscr{E}$'s maximal value) indicates the point at which explanans h fully explains its explanandum e, and $\mathscr{E}(e,h) = -1$ ($\mathscr{E}$'s minimal value) indicates the minimal explanatory power for h relative to e (where h provides a full explanation for e being *false*). $\mathscr{E}(e,h) = 0$ represents the neutral point at which h lacks any explanatory power whatever relative to e.

While we have provided an informal description of the point at which $\mathscr{E}$ should take on its neutral value 0 (when h lacks any explanatory power whatever relative to e), it is still left to us to define this point formally. Given our notion of explanatory power, a complete lack of explanatory power is straightforwardly identified with the scenario in which h does nothing to increase or decrease the degree to which e is surprising. Probabilistically, in such cases, h and e are statistically irrelevant to (independent of) one another:

CA6 (Neutrality**):** For explanatory hypothesis h, $\mathscr{E}(e,h) = 0$ if and only if $Pr(h \wedge e) = Pr(h) \times Pr(e)$.

The final adequacy condition requires that the more h explains e, the less it explains its negation. This requirement is appropriate given that the less surprising (more expected) the truth of e is in light of a hypothesis, the more surprising (less expected) is e's falsity. Corollary 1 and Neutrality provide a further rationale for this condition. Corollary 1 tells us that $\mathscr{E}(e,h)$ should be maximal only if $Pr(e|h) = 1$. Importantly, in such a case, $Pr(\neg e|h) = 0$, and this value corresponds to the point at which this same corollary demands $\mathscr{E}(\neg e,h)$ to be minimal. In other words, given Corollary 1, we see that $\mathscr{E}(e,h)$ takes its maximal value precisely when $\mathscr{E}(\neg e,h)$ takes its minimal value and vice versa. Also, we know that $\mathscr{E}(e,h)$ and $\mathscr{E}(\neg e,h)$ should always equal zero at the same point given that $Pr(h \wedge e) = Pr(h) \times Pr(e)$ if and only if $Pr(h \wedge \neg e) = Pr(h) \times Pr(\neg e)$. The formal condition of adequacy which most naturally sums up all of these points is the following.

CA7 (Symmetry**):** $\mathscr{E}(e,h) = -\mathscr{E}(\neg e,h)$.

These three conditions of adequacy, when added to **CA1-CA4**, conjointly determine a unique measure of explanatory power as stated in the following theorem (Proof in **Appendix 2**).[11]

[11]There is another attractive uniqueness theorem for $\mathscr{E}$. It can be proven that $\mathscr{E}$ is also the only measure that satisfies **CA3**, **CA5**, **CA6**, **CA7**, and Corollary 1, although we do not include this separate proof in this paper.

Theorem 2. *The only measure that satisfies* **CA1-CA7** *is*

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

Remark: Since, for $Pr(h|\neg e) \neq 0$,

$$\frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} = \frac{Pr(h|e)/Pr(h|\neg e) - 1}{Pr(h|e)/Pr(h|\neg e) + 1},$$

it is easy to see that $\mathscr{E}$ is indeed an increasing function of the posterior ratio.

Thus, these conditions provide us with an intuitively grounded, unique measure of explanatory power.[12]

3 Theorems of $\mathscr{E}$

We have proposed the above seven conditions of adequacy as intuitively plausible constraints on a measure of explanatory power. Accordingly, the fact that these conditions are sufficient to determine $\mathscr{E}$ already constitutes a strong argument in this measure's favor. Nonetheless, we proceed in this section to strengthen the case for $\mathscr{E}$ by highlighting some important theorems that follow from adopting this measure. Ultimately, the point of this section is to defend further our assertion that $\mathscr{E}$ is well-behaved in the sense that it gives results that match our clear intuitions about the concept of explanatory power – even in one case where other proposed measures fail to do so.[13]

3.1 Addition of Irrelevant Evidence

Good (1960) and, more recently, McGrew (2003) both explicate h's degree of explanatory power relative to e in terms of the amount of information concerning h provided by e. This results in the following intuitive and simple measure of explanatory power:[14]

[12] $\mathscr{E}$ is closely related to Kemeny and Oppenheim's (1952) measure of "factual support" F. In fact, these two measures are structurally equivalent; however, regarding the interpretation of the measure, $\mathscr{E}(e,h)$ is $F(h,e)$ flip-flopped (h is replaced by e, and e is replaced by h).

[13] Each of the theorems presented in this section can and should be thought of as further conditions of adequacy on any measure of explanatory power. Nonetheless, we choose to present these theorems as separate from the conditions of adequacy presented in Section 2 in order to make explicit which conditions do the work in giving us a unique measure.

[14] Good's measure is meant to improve upon the following measure of explanatory power defined by Popper (1959): $[Pr(e|h) - Pr(e)]/[Pr(e|h) + Pr(e)]$. It should be noted that Popper's measure is ordinally equivalent to Good's in the same sense that $\mathscr{E}$ is ordinally equivalent to the posterior ratio $Pr(h|e)/Pr(h|\neg e)$. Thus, the problem we present here for Good's measure is also a problem for Popper's.

$$I(e,h) = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right]$$

According to this measure, the explanatory power of explanans h must remain constant whenever we add an irrelevant proposition e' to explanandum e (where proposition e' is irrelevant in the sense that it is statistically independent of h in the light of e):

$$\begin{aligned} I(e \wedge e', h) &= \ln\left[\frac{Pr(e \wedge e'|h)}{Pr(e \wedge e')}\right] = \ln\left[\frac{Pr(e'|e \wedge h)Pr(e|h)}{Pr(e'|e)Pr(e)}\right] \\ &= \ln\left[\frac{Pr(e'|e)Pr(e|h)}{Pr(e'|e)Pr(e)}\right] = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right] = I(e,h) \end{aligned}$$

This is, however, a very counterintuitive result. To see this, consider the following simple example: Let e be a general description of the Brownian motion observed in some particles suspended in a particular liquid, and let h be Einstein's atomic explanation of this motion. Of course, h constitutes a lovely explanation of e, and this fact is reflected nicely by measure I:

$$I(e,h) = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right] \gg 0$$

However, take any irrelevant new statement e' and conjoin it to e; for example, let e' be the proposition that the mating season for an American green tree frog takes place from mid-April to mid-August. In this case, measure I judges that Einstein's hypothesis explains Brownian motion to the same extent that it explains Brownian motion *and* this fact about tree frogs. Needless to say, this result is deeply unsettling.

Instead, it seems that, as 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 ($\mathscr{E} = 0$). Thus, if $\mathscr{E}(e,h) > 0$, then this value should *decrease* with the addition of e' to our evidence: $0 < \mathscr{E}(e \wedge e', h) < \mathscr{E}(e,h)$. Similarly, if $\mathscr{E}(e,h) < 0$, then this value should *increase* with the addition of e': $0 > \mathscr{E}(e \wedge e', h) > \mathscr{E}(e,h)$. And finally, if $\mathscr{E}(e,h) = 0$, then this value should *remain constant* at $\mathscr{E}(e \wedge e', h) = 0$. $\mathscr{E}$ gives these general results as shown in the following theorem (proof in **Appendix 3**):

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

- *if* $Pr(e|h) > Pr(e)$*, then* $\mathscr{E}(e,h) > \mathscr{E}(e \wedge e', h) > 0$*,*
- *if* $Pr(e|h) < Pr(e)$*, then* $\mathscr{E}(e,h) < \mathscr{E}(e \wedge e', h) < 0$*, and*
- *if* $Pr(e|h) = Pr(e)$*, then* $\mathscr{E}(e,h) = \mathscr{E}(e \wedge e', h) = 0$*.*

3.2 Addition of Relevant Evidence

Next, we explore whether $\mathscr{E}$ is well-behaved in those circumstances where we strengthen our explanandum by adding to it *relevant* evidence. Consider the case where h has some explanatory power relative to e so that $\mathscr{E}(e,h) > -1$ (i.e., h has any degree of explanatory power relative to e greater than the minimal degree). What should happen to this degree of explanatory power if we gather some new information e' that, in the light of e, we know is explained by h to the *worst* possible degree?

To take a simple example, imagine that police investigators hypothesize that Jones murdered Smith (h) in light of the facts that Jones' fingerprints were found near the dead body and Jones recently had discovered that his wife and Smith were having an affair (e). Now suppose that the investigators discover video footage that proves that Jones was *not* at the scene of the murder on the day and time that it took place (e'). Clearly, h is no longer such a good explanation of our evidence once e' is added; in fact, h now seems to be a maximally poor explanation of $e \wedge e'$ precisely because of the addition of e' (h cannnot possibly explain $e \wedge e'$ because e' rules h out entirely). Thus, in such cases, the explanatory power of h relative to the new collection of evidence $e \wedge e'$ should be less than that relative to the original evidence e; in fact, it should be minimal with the addition of e'. This holds true in terms of $\mathscr{E}$ as shown in the following theorem (proof in **Appendix 4**):

Theorem 4. *If* $\mathscr{E}(e,h) > -1$ *and* $Pr(e'|e \wedge h) = 0$ *(in which case, it also must be true that* $Pr(e'|e) \neq 1$*), then* $\mathscr{E}(e,h) > \mathscr{E}(e \wedge e', h) = -1$*.*

On the other hand, we may ask what intuitively should happen in the same circumstance (adding the condition that h does not have the maximal degree of explanatory power relative to e – i.e., $\mathscr{E}(e,h) < 1$) but where the new information we gain e' is fully explained by h in the light of our evidence e. Let h and e be the same as in the above example, and now imagine that investigators discover video footage that proves that Jones *was* at the scene of the murder on the day and time that it took place (e'). In this case, h becomes an even better explanation of the evidence precisely because of the addition of e' to the evidence. Thus, in such cases, we would expect the explanatory power of h

relative to the new evidence $e \wedge e'$ to be greater than that relative to e alone. Again, $\mathscr{E}$ agrees with our intuition here (proof in **Appendix 4**):

Theorem 5. *If* $0 < Pr(e'|e) < 1$ *and* h *does not already fully explain* e *or its negation (*$0 < Pr(e|h) < 1$*) and* $Pr(e'|e \wedge h) = 1$*, then* $\mathscr{E}(e,h) < \mathscr{E}(e \wedge e', h)$.

While these last two theorems are highly intuitive, they are also quite limited in their applicability. Both theorems require in their antecedent conditions that one's evidence be strengthened with the addition of some e' that is itself either *maximally* or *minimally* explained by h in the light of e. However, our intuitions reach to another class of related examples in which the additional evidence need not be maximally or minimally explained in this way. In situations where h explains e to some positive degree, it is intuitive to think that the addition of any new piece of evidence that is negatively explained by (made more surprising by) h in the light of e will decrease h's degree of explanatory power. Similarly, whenever h has some negative degree of explanatory power relative to e, it is plausible to think that the addition of any new piece of evidence that is positively explained by (made less surprising by) h in the light of e will increase h's degree of explanatory power. These intuitions are captured in the following theorem of $\mathscr{E}$ (proof in **Appendix 4**):

Theorem 6. *If* $\mathscr{E}(e,h) > 0$*, then if* $Pr(e'|e \wedge h) < Pr(e'|e)$*, then* $\mathscr{E}(e \wedge e', h) < \mathscr{E}(e,h)$*. On the other hand, if* $\mathscr{E}(e,h) < 0$*, then if* $Pr(e'|e \wedge h) > Pr(e'|e)$*, then* $\mathscr{E}(e \wedge e', h) > \mathscr{E}(e,h)$.

4 Conclusions

Above, we have shown the following: First, $\mathscr{E}$ is a member of the specific family of ordinally equivalent measures that satisfy our first four adequacy conditions. Moreover, among the measures included in this class, $\mathscr{E}$ itself *uniquely* satisfies the additional conditions **CA5-CA7**. The theorems presented in the last section strengthen the case for $\mathscr{E}$ by showing that this measure does indeed seem to correspond well and quite generally to many of our clear explanatory intuitions. In light of all of this, we argue that $\mathscr{E}$ is manifestly an intuitively appealing formal account of explanatory power.

The acceptance of $\mathscr{E}$ opens the door to a wide variety of potentially fruitful, intriguing further questions for research. Here, we limit ourselves to describing very briefly two of these projects which seem to us to be particularly fascinating and manageable with $\mathscr{E}$ in hand.

First, measure $\mathscr{E}$ makes questions pertaining to the normativity of explanatory considerations much more tractable, at least from a Bayesian perspective.

Given this probabilistic rendering of the concept of explanatory power, one has a new ability to ask and attempt to answer questions such as, "Does the ability of a hypothesis to explain some known fact itself constitute reason in that hypothesis's favor in any sense?" or, relatedly, "Is there any necessary sense in which explanatory power is tied to the probability of an hypothesis?" Such questions call out for more formal work in terms of $\mathscr{E}$ attempting to show whether, and how closely, $\mathscr{E}(e,h)$ might be related to $Pr(h|e)$. This further work would have important bearing on debates over the general normativity of explanatory power; it would also potentially lend much insight into discussions of Inference to the Best Explanation and its vices or virtues.

Second, we have presented and defended $\mathscr{E}$ here as an accurate *normative* account of explanatory power in the following sense: In the wide space of cases where our conditions of adequacy are rationally compelling and intuitively applicable, one *ought* to think of explanatory power in accord with the results of $\mathscr{E}$. However, one may wonder whether people actually have explanatory intuitions that accord with this normative measure. With $\mathscr{E}$ in hand, this question becomes quite susceptible to further study. In effect, the question is whether $\mathscr{E}$ is, in addition to being an accurate normative account of explanatory power, a good *predictor* of people's actual judgments of the same. This question is, of course, an empirical one and thus requires an empirical study into the degree of fit between human judgments and theoretical results provided by $\mathscr{E}$. Such a study could provide important insights both for the psychology of human reasoning and for the philosophy of explanation.[15]

[15]This second research project is, in fact, now underway. For a description and report of the first empirical study investigating the descriptive merits of $\mathscr{E}$ (and other candidate measures of explanatory power), see (Schupbach, 2010).

Appendix 1. Proof of **Theorem 1** and **Corollary 1**.

Theorem 1. *All measures of explanatory power satisfying* **CA1-CA4** *are monotonically increasing functions of the posterior ratio* $Pr(h|e)/Pr(h|\neg e)$.

Proof. $Pr(h|e)$, $Pr(h|\neg e)$ and $Pr(e)$ jointly determine the probability distribution of the pair (e,h); so we can represent $\mathscr{E}$ as a function of these values: there is a $g:[0,1]^3 \to \mathbb{R}$ such that $\mathscr{E}(e,h) = g(Pr(e), Pr(h|e), Pr(h|\neg e))$.

First, note that whenever the assumptions of **CA3** are satisfied (i.e., whenever h_2 is independent of all e, h_1 and $e \wedge h_1$), the following equalities hold:

$$\begin{aligned} Pr(h_1 \wedge h_2|e) &= Pr(h_2|h_1 \wedge e)Pr(h_1|e) = Pr(h_2)Pr(h_1|e) \\ Pr(h_1 \wedge h_2|\neg e) &= \frac{Pr(h_1 \wedge h_2 \wedge \neg e)}{Pr(\neg e)} = \frac{Pr(h_1 \wedge h_2) - Pr(h_1 \wedge h_2 \wedge e)}{Pr(\neg e)} \\ &= Pr(h_2)\frac{Pr(h_1) - Pr(h_1 \wedge e)}{Pr(\neg e)} = Pr(h_2)Pr(h_1|\neg e). \end{aligned} \tag{1}$$

Now, for all values of $c, x, y, z \in (0,1)$, we can choose propositions e, h_1 and h_2 and probability distributions over these such that the independence assumptions of **CA3** are satisfied and $c = Pr(h_2)$, $x = Pr(e)$, $y = Pr(h_1|e)$, and $z = Pr(h_1|\neg e)$. Due to **CA1**, we can always find such propositions and distributions so long as $\mathscr{E}$ is applicable. The above equations then imply that $Pr(h_1 \wedge h_2|e) = cy$ and $Pr(h_1 \wedge h_2|\neg e) = cz$. Applying **CA3** ($\mathscr{E}(e,h_1) = \mathscr{E}(e,h_1 \wedge h_2)$) yields the general fact that

$$g(x,y,z) = g(x,cy,cz). \tag{2}$$

Consider now the case that $\neg h$ entails e; i.e., $Pr(e|\neg h) = Pr(h|\neg e) = 1$. Assume that $g(\cdot,\cdot,1)$ could be written as a function of $Pr(e)$ alone. Accordingly, there would be a function $h:[0,1] \to \mathbb{R}$ such that

$$g(x,y,1) = h(x). \tag{3}$$

If we choose $y = Pr(h|e) < Pr(h|\neg e) = z$, it follows from equations (2) and (3) that

$$g(x,y,z) = g(x,y/z,1) = h(x). \tag{4}$$

In other words, g (and $\mathscr{E}$) would then be constant on the triangle $\{y < z\} = \{Pr(h|e) < Pr(h|\neg e)\}$ for any fixed $x = Pr(e)$. Now, since g is an analytic function (due to **CA1**), its restriction $g(x,\cdot,\cdot)$ (for fixed x) must be analytic as well. This entails in particular that if $g(x,\cdot,\cdot)$ is constant on some nonempty open set $S \subset R^2$, then it is constant everywhere:

1. All derivatives of a locally constant function vanish in that environment (Theorem of Calculus).

2. We write, by **CA1**, $g(x,\cdot,\cdot)$ as a Taylor series expanded around a fixed point $(y^*,z^*)\in S=\{y<z\}$:

$$g(x,y,z)=\sum_{j=0}^{\infty}\left[\frac{1}{j!}\left((y-y^*)\frac{\partial}{\partial y}+(z-z^*)\frac{\partial}{\partial z}\right)^j g(x,y^*,z^*)\right]_{y=y^*,z=z^*}.$$

   Since all derivatives of $g(x,\cdot,\cdot)$ in the set $S=\{y<z\}$ are zero, all terms of the Taylor series, except the first one ($=g(x,y^*,z^*)$) vanish.

Thus, $g(x,\cdot,\cdot)$ must be constant everywhere. But this would violate the statistical relevance condition **CA2** since g (and $\mathscr{E}$) would then depend on $Pr(e)$ alone and not be sensitive to any form of statistical relevance between e and h.

Thus, whenever $\neg h$ entails e, $g(\cdot,\cdot,1)$ either depends on its second argument alone, or on both arguments. The latter case implies that there must be pairs (e,h) and (e',h') with $Pr(h|e)=Pr(h'|e')$ such that

$$g(Pr(e),Pr(h|e),1)\neq g(Pr(e'),Pr(h'|e'),1). \tag{5}$$

Note that if $Pr(e|\neg h)=1$, we obtain

$$\begin{aligned} Pr(e) &= Pr(e|h)Pr(h)+Pr(e|\neg h)Pr(\neg h) = Pr(h|e)Pr(e)+(1-Pr(h)) \\ &= \frac{1-Pr(h)}{1-Pr(h|e)}, \end{aligned} \tag{6}$$

and so we can write $Pr(e)$ as a function of $Pr(h)$ and $Pr(h|e)$.

Combining (5) and (6), and keeping in mind that g cannot depend on $Pr(e)$ alone, we obtain that there are pairs (e,h) and (e',h') such that

$$g\left(\frac{1-Pr(h)}{1-Pr(h|e)},Pr(h|e),1\right)\neq g\left(\frac{1-Pr(h')}{1-Pr(h|e)},Pr(h|e),1\right).$$

This can only be the case if the prior probability ($Pr(h)$ and $Pr(h')$ respectively) has an impact on the value of g (and thus on $\mathscr{E}$), in contradiction with **CA4**. Thus, equality in (5) holds whenever $Pr(h|e)=Pr(h'|e')$. Hence, $g(\cdot,\cdot,1)$ cannot depend on both arguments, and it can be written as a function of its second argument alone.

Thus, for any $Pr(h|e)<Pr(h|\neg e)$, there must be a $g':[0,1]^2\to\mathbb{R}$ such that

$$\begin{aligned} \mathscr{E}(e,h) &= g(Pr(e),Pr(h|e),Pr(h|\neg e))=g(Pr(e),Pr(h|e)/Pr(h|\neg e),1) \\ &= g'(Pr(h|e)/Pr(h|\neg e),1). \end{aligned}$$

This establishes that $\mathscr{E}$ is a function of the posterior ratio if h and e are negatively relevant to each other. By applying analyticity of $\mathscr{E}$ once more, we see that $\mathscr{E}$ is a function of the posterior ratio $Pr(h|e)/Pr(h|\neg e)$ in its entire domain (i.e. also if e and h are positively relevant to each other or independent).

Finally, **CA2** implies that this function must be monotonically increasing, since otherwise, explanatory power would not increase with statistical relevance (of which the posterior probability is a measure). Evidently, any such function satisfies **CA1-CA4**.

□

Corollary 1: *$\mathscr{E}(e,h)$ takes maximal value if and only if h entails e, and minimal value if and only if h implies $\neg e$.*

Proof. Since $\mathscr{E}$ is an increasing function of the posterior ratio $Pr(h|e)/Pr(h|\neg e)$, $\mathscr{E}(e,h)$ is maximal if and only if $Pr(h|\neg e) = 0$. Due to the regularity of $Pr(\cdot)$, this is the case of and only if $\neg e$ entails $\neg h$, in other words, if and only if h entails e. The case of minimality is proven analogously.

□

Appendix 2. Proof of **Theorem 2** (Uniqueness of $\mathscr{E}$).

Theorem 2. *The only measure that satisfies* **CA1-CA7** *(i.e., the only measure of explanatory power) is*

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

Let $x = Pr(e \wedge h)$, $y = Pr(e \wedge \neg h)$, $z = Pr(\neg e \wedge h)$ and $t = Pr(\neg e \wedge \neg h)$ with $x + y + z + t = 1$. Write $\mathscr{E}(e,h) = f(x,y,z,t)$ (by **CA5**).

Lemma 1. *There is no normalized function $f(x,y,z,t)$ of degree 1 that satisfies our desiderata* **CA1-CA7**.

Proof. If there were such a function, the numerator would have the form $ax + by + cz + dt$. If e and h are independent, the numerator must vanish, by means of CA6. In other words, for those values of (x,y,z,t), we demand $ax + by + cz + dt = 0$. Below, we list four different realizations of (x,y,z,t) that make e and h independent, namely $(1/2, 1/4, 1/6, 1/12)$, $(1/2, 1/3, 1/10, 1/15)$, $(1/2, 3/8, 1/14, 3/56)$, and $(1/4, 1/4, 1/4, 1/4)$. Since these vectors are linearly independent (i.e. their span has dimension 4), it must be the case that $a = b = c = d = 0$. Hence there is no such function of degree 1.

□

Lemma 2. **CA3** *entails that for any value of $\beta \in (0,1)$,*

$$f(\beta x, y + (1-\beta)x, \beta z, t + (1-\beta)z) = f(x,y,z,t). \tag{7}$$

Proof. For any x, y, z, t, we choose e, h_1 such that

$$\begin{aligned} x &= Pr(e \wedge h_1) & y &= Pr(e \wedge \neg h_1) \\ z &= Pr(\neg e \wedge h_1) & t &= Pr(\neg e \wedge \neg h_1). \end{aligned}$$

Moreover, we choose a h_2 such that the antecedent conditions of **CA3** are satisfied, and we let $\beta = Pr(h_2)$. Applying the independencies between h_2, e and h_1 and recalling (1), we obtain

$$\begin{aligned} \beta x &= Pr(h_2)Pr(e \wedge h_1) = Pr(h_2)Pr(e)Pr(h_1|e) \\ &= Pr(e)Pr(h_1 \wedge h_2|e) = Pr(e \wedge h_1 \wedge h_2), \end{aligned}$$

and similarly

$$\beta z = Pr(\neg e \wedge (h_1 \wedge h_2)) \qquad y + (1-\beta)z = Pr(\neg e \wedge \neg(h_1 \wedge h_2))$$
$$y + (1-\beta)x = Pr(e \wedge \neg(h_1 \wedge h_2)).$$

Making use of these equations, we see directly that **CA3** – i.e. $\mathscr{E}(e, h_1) = \mathscr{E}(e, h_1 \wedge h_2)$ – implies equation (7). □

Proof of Theorem 2 (Uniqueness of $\mathscr{E}$). Lemma 1 shows that there is no normalized function $f(x, y, z, t)$ of degree 1 that satisfies our desiderata. Our proof is constructive: we show that there is exactly one such function of degree 2, and then we are done, due to the formal requirements set out in **CA5**. By **CA5**, we look for a function of the form

$$f(x, y, z, t) = \frac{ax^2 + bxy + cy^2 + dxz + eyz + gz^2 + ixt + jyt + rzt + st^2}{\bar{a}x^2 + \bar{b}xy + \bar{c}y^2 + \bar{d}xz + \bar{e}yz + \bar{g}z^2 + \bar{i}xt + \bar{j}yt + \bar{r}zt + \bar{s}t^2} \tag{8}$$

We begin by investigating the numerator.[16] **CA6** tells us that it has to be zero if $Pr(e \wedge h) = Pr(e)Pr(h)$, in other words, if

$$x = (x + y)(x + z). \tag{9}$$

Making use of $x + y + z + t = 1$, we conclude that this is the case if and only if $xt - yz = 0$:

$$\begin{aligned} xt - yz &= x(1 - x - y - z) - yz \\ &= x - x^2 - xy - xz - yz \\ &= x - (x + y)(x + z) \end{aligned}$$

The only way to satisfy the constraint (9) is to set $e = -i$, and to set all other coefficients in the numerator to zero. All other choices of coefficients don't work since the dependencies are non-linear. Hence, f becomes

$$f(x, y, z, t) = \frac{i(xt - yz)}{\bar{a}x^2 + \bar{b}xy + \bar{c}y^2 + \bar{d}xz + \bar{e}yz + \bar{g}z^2 + \bar{i}xt + \bar{j}yt + \bar{r}zt + \bar{s}t^2}$$

[16] The general method of our proof bears resemblance to Kemeny and Oppenheim's Theorem 27 in their (1952). However, we would like to point out two crucial differences. First, we use more parsimonious assumptions, and we work in a different – non-Carnapian – framework. Second, their proof contains invalid steps, for instance, they derive $d = 0$ by means of symmetry (**CA7**) alone. (Take the counterexample $f = (xy - yz + xz - z^2)/(xy + yz + xz - z^2)$ which even satisfies **Corollary 1**.) Hence, our proof is truly original.

Now, we make use of **Corollary 1** and **CA7** in order to tackle the coefficients in the denominator. **Corollary 1** (Maximality) entails that $f = 1$ if $z = 0$, and **CA7** (Symmetry) is equivalent to

$$f(x,y,z,t) = -f(z,t,x,y). \tag{10}$$

First, applying **Corollary 1** yields $1 = f(x,0,0,t) = ixt/(\bar{a}x^2 + \bar{i}xt + \bar{s}t^2)$, and by a comparison of coefficients, we get $\bar{a} = \bar{s} = 0$ and $\bar{i} = i$. Similarly, we obtain $\bar{c} = \bar{g} = 0$ and $\bar{e} = i$ from $1 = f(x,0,0,t) = -f(0,t,x,0) = ixt/(\bar{c}t^2 + \bar{e}xt + \bar{g}x^2)$, combining **Corollary 1** with **CA7**, i.e. equation (10).

Now, f has the form

$$f(x,y,z,t) = \frac{i(xt - yz)}{\bar{b}xy + \bar{d}xz + i(xt + yz) + \bar{j}yt + \bar{r}zt}.$$

Assume now that $\bar{j} \neq 0$. Let $x, z \to 0$. We know by **Corollary 1** that in this case, $f \to 1$. Since the numerator vanishes, the denominator must vanish too, but by $\bar{j} \neq 0$ it stays bounded away from zero, leading to a contradiction ($f \to 0$). Hence $\bar{j} = 0$. In a similar vein, we can argue for $\bar{b} = 0$ by letting $z, t \to 0$ and for $\bar{r} = 0$ by letting $x, y \to 0$ (making use of (10) again: $-1 = f(0,0,z,t)$).

Thus, f can be written as

$$f(x,y,z,t) \;=\; \frac{i(xt - yz)}{\bar{d}xz + \bar{i}(xt + yz)} = \frac{(xt - yz)}{(xt + yz) + \alpha xz}, \tag{11}$$

by letting $\alpha = \bar{d}/i$.

It remains to make use of **CA3** in order to fix the value of α. Set $\beta = 1/2$ in (7) and make use of $f(x,y,z,t) = f(\beta x, (1-\beta)x + y, \beta z, (1-\beta)z + t)$ (**Lemma 2**) and the restrictions on f captured in (11). By making use of (7), we obtain the general constraint

$$\begin{aligned} \frac{xt - yz}{xt + yz + \alpha xz} &= \frac{x(z/2 + t) - z(x/2 + y)}{x(z/2 + t) + z(x/2 + y) + \alpha xz/2} \\ &= \frac{xt - yz}{xt + yz + xz(2 + \alpha)/2} \end{aligned} \tag{12}$$

For (12) to be true in general, we have to demand that $\alpha = 1 + \alpha/2$ which implies that $\alpha = 2$. Hence,

$$f(x,y,z,t) = \frac{xt - yz}{xt + yz + 2xz} = \frac{x(t+z) - z(x+y)}{x(t+z) + z(x+y)}$$

implying

$$\begin{aligned} \mathscr{E}(e,h) &= \frac{Pr(e \wedge h)Pr(\neg e) - Pr(\neg e \wedge h)Pr(e)}{Pr(e \wedge h)Pr(\neg e) + Pr(\neg e \wedge h)Pr(e)} \\ &= \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} \end{aligned} \tag{13}$$

which is the unique function satisfying all our desiderata.

□

Appendix 3. Proof of **Theorem 3**.

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

- *if $Pr(e|h) > Pr(e)$, then $\mathscr{E}(e,h) > \mathscr{E}(e \wedge e', h) > 0$,*
- *if $Pr(e|h) < Pr(e)$, then $\mathscr{E}(e,h) < \mathscr{E}(e \wedge e', h) < 0$, and*
- *if $Pr(e|h) = Pr(e)$, then $\mathscr{E}(e,h) = \mathscr{E}(e \wedge e', h) = 0$.*

Proof. Since $\mathscr{E}(e,h)$ and the posterior ratio $r(e,h) = Pr(h|e)/Pr(h|\neg e)$ are ordinally equivalent, we can focus our analysis on that quantity:

$$\begin{aligned}
\frac{r(e,h)}{r(e \wedge e', h)} &= \frac{Pr(h|e)}{Pr(h|\neg e)} \cdot \frac{Pr(h|\neg(e \wedge e'))}{Pr(h|e \wedge e')} \\
&= \frac{1 - Pr(e)}{Pr(e)} \cdot \frac{Pr(e|h)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e'|e \wedge h) Pr(e|h)}{Pr(e'|e \wedge h) Pr(e|h)} \cdot \frac{Pr(e \wedge e')}{1 - Pr(e \wedge e')} \\
&= \frac{1 - Pr(e)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e'|e) Pr(e|h)}{1 - Pr(e) Pr(e'|e)} \\
&= \frac{1 + Pr(e) Pr(e'|e) Pr(e|h) - (Pr(e) + Pr(e|h) Pr(e'|e))}{1 + Pr(e) Pr(e'|e) Pr(e|h) - (Pr(e|h) + Pr(e) Pr(e'|e))} \qquad (14)
\end{aligned}$$

This quantity is greater than one if and only if the numerator exceeds the denominator, i.e. iff

$$\begin{aligned}
0 &< (Pr(e|h) + Pr(e) Pr(e'|e)) - (Pr(e) + Pr(e|h) Pr(e'|e)) \\
&= Pr(e|h)(1 - Pr(e'|e)) - Pr(e)(1 - Pr(e'|e)) \\
&= (Pr(e|h) - Pr(e))(1 - Pr(e'|e)) \qquad (15)
\end{aligned}$$

which is satisfied if and only if $Pr(e|h) > Pr(e)$, and not satisfied otherwise. Thus, $r(e,h) > r(e \wedge e', h)$ (and $\mathscr{E}(e,h) > \mathscr{E}(e \wedge e', h)$) if and only if $Pr(e|h) > Pr(e)$. The other two cases follow directly from (15).

It remains to show that $\mathscr{E}(e,h)$ and $\mathscr{E}(e \wedge e', h)$ always have the same sign. This follows from the fact that

$$\frac{Pr(e \wedge e'|h)}{Pr(e \wedge e')} = \frac{Pr(e'|e \wedge h)\, Pr(e|h)}{Pr(e|e')\, Pr(e)} = \frac{Pr(e|h)}{Pr(e)}.$$

Thus, h is positively relevant to e if and only if it is positively relevant to $(e \wedge e')$. By **CA2** and **CA6**, this implies that $\mathscr{E}(e \wedge e', h) > 0$ if and only if $\mathscr{E}(e, h) > 0$, and vice versa for negative relevance.

□

Appendix 4. Proofs of **Theorem 4-6**.

Theorem 4. *If $\mathscr{E}(e,h) > -1$ and $Pr(e'|e \wedge h) = 0$ – in which case, it also must be true that $Pr(e'|e) \neq 1$ – then $\mathscr{E}(e,h) > \mathscr{E}(e \wedge e', h) = -1$.*

Proof. Under the assumptions of Theorem 1, by application of Bayes's Theorem,

$$Pr(h|e \wedge e') = \frac{Pr(h)\,Pr(e \wedge e'|h)}{Pr(e \wedge e')} = \frac{Pr(h)\,Pr(e'|e \wedge h)\,Pr(e|h)}{Pr(e \wedge e')} = 0.$$

Thus

$$\mathscr{E}(e \wedge e'|h) = -1 < \mathscr{E}(e,h).$$

□

Theorem 5. *If $0 < Pr(e'|e) < 1$ and h does not already fully explain e or its negation – i.e., $0 < Pr(e|h) < 1$ – and $Pr(e'|e \wedge h) = 1$, then $\mathscr{E}(e,h) < \mathscr{E}(e \wedge e', h)$.*

Proof. Note first that

$$Pr(e \wedge e'|h) = Pr(e'|e \wedge h)\,Pr(e|h) = Pr(e|h). \tag{16}$$

Analogous to Theorem 3, we prove this theorem by comparing the posterior ratios $r(e,h)$ and $r(e \wedge e', h)$, and applying equation (16):

$$\begin{aligned}
\frac{r(e,h)}{r(e \wedge e', h)} &= \frac{Pr(h|e)}{Pr(h|\neg e)} \cdot \frac{Pr(h|\neg(e \wedge e'))}{Pr(h|e \wedge e')} \\
&= \frac{1 - Pr(e)}{Pr(e)} \cdot \frac{Pr(e|h)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e \wedge e'|h)}{Pr(e \wedge e'|h)} \cdot \frac{Pr(e \wedge e')}{1 - Pr(e \wedge e')} \\
&= \frac{1 - Pr(e)}{Pr(e)} \cdot \frac{Pr(e \wedge e')}{1 - Pr(e \wedge e')} \\
&= \frac{Pr(e \wedge e') - Pr(e)Pr(e \wedge e')}{Pr(e) - Pr(e)Pr(e \wedge e')} \\
&< 1,
\end{aligned}$$

since, by assumption, $Pr(e \wedge e') = Pr(e)Pr(e'|e) < Pr(e)$. This implies that $\mathscr{E}(e,h) < \mathscr{E}(e \wedge e', h)$.

□

Theorem 6. *If $\mathscr{E}(e,h) > 0$, then if $Pr(e'|e \wedge h) < Pr(e'|e)$, then $\mathscr{E}(e \wedge e', h) < \mathscr{E}(e,h)$. On the other hand, if $\mathscr{E}(e,h) < 0$, then if $Pr(e'|e \wedge h) > Pr(e'|e)$, then $\mathscr{E}(e \wedge e', h) > \mathscr{E}(e,h)$.*

Proof. First, we note that if $Pr(e'|e \wedge h) < Pr(e'|e)$, then also $Pr(e \wedge e'|h) = Pr(e'|e \wedge h)Pr(e|h) < Pr(e'|e)Pr(e|h)$. Then we apply the same approach as in the previous proofs:

$$\begin{aligned}
\frac{r(e,h)}{r(e \wedge e', h)} &= \frac{1 - Pr(e)}{Pr(e)} \cdot \frac{Pr(e|h)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e \wedge e'|h)}{Pr(e \wedge e'|h)} \cdot \frac{Pr(e \wedge e')}{1 - Pr(e \wedge e')} \\
&> \frac{1 - Pr(e)}{Pr(e)} \cdot \frac{Pr(e|h)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e'|e)Pr(e|h)}{Pr(e'|e)Pr(e|h)} \cdot \frac{Pr(e'|e)Pr(e)}{1 - Pr(e'|e)Pr(e)} \\
&= \frac{1 - Pr(e)}{1 - Pr(e|h)} \cdot \frac{1 - Pr(e|h)Pr(e'|e)}{1 - Pr(e'|e)Pr(e)} \\
&= \frac{1 + Pr(e)Pr(e'|e)Pr(e|h) - (Pr(e) + Pr(e|h)Pr(e'|e))}{1 + Pr(e)Pr(e'|e)Pr(e|h) - (Pr(e|h) + Pr(e)Pr(e'|e))}
\end{aligned}$$

This is exactly the term in the last line of (14). We have already shown in the proof of Theorem 3 that this quantity is greater than 1 if and only if $Pr(e|h) > Pr(e)$, i.e. if $\mathscr{E}(e,h) > 0$. This suffices to prove the first half of Theorem 6. The reverse case is proved in exactly the same way.

□

References

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

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy 71*(1), 5–19.

Glymour, C. (1980). Explanations, Tests, Unity and Necessity. *Nous 14*, 31–49.

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.

Hempel, C. G. (1965). Aspects of Scientific Explanation. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pp. 331–496. New York: Free Press.

Hempel, C. G. and P. Oppenheim (1948, April). Studies in the Logic of Explanation. *Philosophy of Science 15*(2), 135–175.

Jeffrey, R. (1969). Statistical Explanation versus Statistical Inference. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel*, pp. 104–113. Dordrecht, Holland: D. Reidel.

Kemeny, J. G. and P. Oppenheim (1952). Degree of Factual Support. *Philosophy of Science 19*, 307–324.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher and W. Salmon (Eds.), *Scientific Explanation*, pp. 410–505. Minneapolis: University of Minnesota Press.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). New York, NY: Routledge.

Machamer, P., L. Darden, and C. F. Craver (2000, March). Thinking about Mechanisms. *Philosophy of Science 67*(1), 1–25.

McGrew, T. (2003). Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science 54*, 553–567.

Peirce, C. S. (1931-1935). *The Collected Papers of Charles Sanders Peirce*, Volume I-VI. Cambridge, Mass: Harvard University Press.

Peirce, C. S. (1958). *The Collected Papers of Charles Sanders Peirce*, Volume VII-VIII. Cambridge, Mass: Harvard University Press.

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

Salmon, W. C. (1971). Statistical Explanation. In W. C. Salmon (Ed.), *Statistical Explanation and Statistical Relevance*, pp. 29–87. Pittsburgh: University of Pittsburgh Press.

Schupbach, J. N. (Forthcoming, 2010). Comparing Probabilistic Measures of Explanatory Power. *Philosophy of Science*.

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

Woodward, J. (Spring 2009). Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.