To appear in *Analysis*

Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory

WILLIAM ROCHE AND ELLIOTT SOBER

If a hypothesis is explanatory, is that evidence that the hypothesis is true? For example, we are thinking of two propositions, H and O, and we tell you this:

(E) If H and O were true, H would explain O.[1]

Is E evidence for H? Bayesian confirmation theory takes this question to be asking whether E raises the probability of H:

(1) $\Pr(H \mid E) > \Pr(H)$.

This inequality will not be plausible if you know from the get-go that O is false. And if you have no clue as to whether O is true, you should not regard E as evidence for H. A better version of our question is whether the conjunction of O and E confirms H:

(2) $\Pr(H \mid O\&E) > \Pr(H)$.

[1] Strictly speaking, (E) is distinct from the claim that if H and O were true, H would provide a 'lovely' or 'satisfying' explanation of O. But what we say about the former claim can also be said *mutatis mutandis* about the latter claim.

But even this way of putting the question isn't quite right. Perhaps this inequality is true whenever O by itself confirms H; in that circumstance, maybe (2) is true simply because E is evidentially irrelevant to H, so adding E to O does no harm. Our question really should be: *does E add anything to O's confirmation of H*? What we really want to know, when $\Pr(H \mid O) > \Pr(H)$, is whether

$$\Pr(H \mid O\&E) > \Pr(H \mid O).$$

Our thesis is that this inequality is false. What is true instead is this:

$$(3)\quad \Pr(H \mid O\&E) = \Pr(H \mid O).$$

This equality says that the observation O *screens-off* E from H; according to proposition (3), the explanatoriness of H is evidentially idle, once the truth of O is taken into account. If you already know that O is true and you have computed $\Pr(H \mid O)$, learning E does not change how confident you should be in H. If we 'relocate' proposition O, shifting it from the conditioning proposition that it is in (3) and making it part of the probability function itself, (3) becomes

$$(4)\quad \Pr_O(H \mid E) = \Pr_O(H).$$

This equality says that E is confirmationally irrelevant to H within the probability function $\Pr_O(\text{-})$. This judgement of irrelevance does not depend on which Bayesian measure of degree of confirmation one adopts (Fitelson 1999).

We will argue for (3) by describing an example. Scientists began studying the relationship of smoking cigarettes and lung cancer by assembling frequency data. They observed that people who smoke more cigarettes get lung cancer more frequently than people who smoke fewer. These observations convinced them that a probabilistic inequality is true:

> Pr(S will get lung cancer $\mid$ S has smoked i cigarettes to date) > Pr(S will get lung cancer $\mid$ S has smoked j cigarettes to date), for all $i > j$.

This inequality says that smoking and cancer are correlated; it leaves open whether the one causes the other. We'll return to this causal question in a moment, but first we want to consider statements that assign a value to a conditional probability in which the placement of smoking and lung cancer are reversed from the probabilities that figure in this last inequality. For example, consider this one:

(5) Pr(*S* smoked at least 10,000 cigarettes before age 50 | *S* got lung cancer after age 50) = *c*.

We will assume that a good estimate of *c* can be found by observing a large group of individuals who contracted lung cancer after age 50 and then seeing how much they had smoked before they reached that age. We will also assume that the probability in (5) is greater than the unconditional probability Pr(*S* smoked at least 10,000 cigarettes before age 50) – lung cancer later in life raises the probability that one was a heavy smoker earlier.

The slogan that 'correlation isn't the same as causation' was of great importance to the investigation of how smoking cigarettes and lung cancer are related. Eventually the hypothesis that smoking causes lung cancer won out, but along the way the distinguished statistician and population geneticist R. A. Fisher (1959) constructed an alternative hypothesis. Fisher's hypothesis was that smoking and lung cancer are joint effects of a common cause; there is a gene that gives you a yen to smoke cigarettes and also causes you to develop lung cancer. Fisher's hypothesis is a possible explanation of the correlation of smoking and lung cancer. To put this dispute between causal and noncausal theories of the correlation between smoking and lung cancer to work, we will assume that causation and explanation are related as follows: if the causal hypothesis is true, then a person's being a heavy smoker would explain why he or she gets lung cancer, whereas if Fisher's hypothesis is true, then a person's being a heavy smoker would not explain why he or she gets lung cancer.

If you deny the screening-off claim expressed in (3) and (4) and hold that explanatoriness is evidentially relevant, you are obliged to endorse the following inequality:

(6) Pr(*S* smoked at least 10,000 cigarettes before age 50 | *S* got lung cancer after age 50 & if *S* smoked at least 10,000 cigarettes before age 50 and *S* got lung cancer subsequently, then the smoking would explain the lung cancer) > Pr(*S* smoked at least 10,000 cigarettes before age 50 | *S* got lung cancer after age 50).

Propositions (5) and (6) are not logically incompatible. However, it is curious that (5) is supported by the frequency data we mentioned, but (6) is not. The frequency of heavy smokers among people who subsequently get lung cancer provides a good estimate of the value of c in (5), and adding the claim that heavy smoking is explanatory doesn't change what that estimate should be. In short, (6) is false; what is true is an equality.

It may be objected that data from a finite sample can be misleading and that this leaves room for asserting inequality (6), the idea being that the observed frequency underestimates the true probability, which the fact about explanatoriness corrects. Our reply is that the problem persists as sample size is increased.

Our interpretation of this example does not mean that the causal hypothesis and Fisher's alternative to it are evidentially indistinguishable. They are not. If smoking causes cancer, then getting people to stop smoking should reduce the incidence of the cancer, but if smoking and lung cancer are joint effects of a common genetic cause, then intervening on people's smoking habits should leave cancer rates unchanged.

This example connects with a wider issue. Explanatory asymmetries are a staple in the literature on explanation; think of Bromberger's (1966) example of the flagpole and the shadow. In general, explanation is asymmetric: if X explains Y, then it does not follow that Y explains X. In contrast, there is no such asymmetry in confirmation. On the contrary, what Bayesianism enshrines is *symmetry*: if X confirms Y, then Y confirms X. This follows from the Bayesian definition of confirmation, since

$$\Pr(Y \mid X) > \Pr(Y) \text{ if and only if } \Pr(X \mid Y) > \Pr(X)$$

is a consequence of Bayes's Theorem. If the explanation relation is asymmetric and the evidence relation is symmetric, it is no surprise that evidential relations are sometimes indifferent to explanatory relations. If smoking is evidence for lung cancer, then lung cancer is evidence for smoking, and it makes no difference that smoking explains lung cancer but lung cancer does not explain smoking. To make this vivid, consider the following equality:

(7) $\Pr(S$ smoked cigarettes earlier in life $\mid S$ gets lung cancer later$) = \Pr(S$ gets lung cancer later in life $\mid S$ smoked cigarettes earlier$)$.

There is no a priori reason why this equality should be true, but suppose that frequency data amply support it. If explanatoriness were evidentially relevant, the following inequality should be true:

(8) Pr(*S* smoked cigarettes earlier in life │ *S* gets lung cancer later & earlier smoking would explain later lung cancer) > Pr(*S* gets lung cancer later in life │ *S* smoked cigarettes earlier & later lung cancer would not explain earlier smoking).

We suggest that if (7) is supported by frequency data, then (8) will not be.

We have railed against the thesis that

Pr(*Y*│*X* & if *X* and *Y* were true, then *Y* would explain *X*) > Pr(*Y*│*X*),

but there is a counterpart thesis that we also want to criticize. It says that

Pr(*Y*│*X* & if *X* and *Y* were true, then *X* would explain *Y*) > Pr(*Y*│*X*).

Here again, we think that what is true is an equality, and smoking and cancer again furnish an illustrative example. We hold that

Pr(*S* gets cancer later in life │ *S* smoked cigarettes earlier & if *S* smoked cigarettes earlier and got lung cancer later, then *S*'s smoking would explain *S*'s getting lung cancer) = Pr(*S* gets cancer later in life │ *S* smoked cigarettes earlier in life).

A good estimate of the probability on the right is furnished by frequency data; the same estimate is a good one for the probability on the left.

In the above examples concerning smoking and cancer, we used frequency data to estimate the value of a probability and then argued that these estimates are not changed when facts about explanatoriness are taken into account. It might be objected that Bayesianism makes room for the idea that the observed frequency is not always the best estimate of a probability. For example, suppose you toss a coin 1000 times and observe that 503 of the tosses landed heads. Suppose you know that almost all coins are fair (meaning that their probability of landing heads when tossed is $p = 0.5$) but that a very small number of coins are heavily biased in favor of heads (with $p = 0.9$). This information will lead you to infer that the coin you tossed has $p = 0.5$. Our reply is to agree that prior 'theoretical' information can influence one's estimate of a probability; observed frequencies are not the only source of information. However, notice in this example that the concept of explanatoriness plays no role.

Our thesis is not that explanatory information never conveys confirmation. For example, in Hempel's (1965) deductive-nomological theory of explanation, H is the *explanans* proposition in a DN explanation of O only if H entails O. We know from Bayes's theorem that if H entails O, then O cannot disconfirm H; O will either leave H's probability unchanged, or O will raise H's probability. So the DN information tells us something about confirmation. However, what is doing the work here is the formal relation of the two propositions H and O. It is just because H entails O that the confirmational fact falls into place; whether H if true would explain O is as it may be. Another feature of Hempel's theory is that if H explains O, then H is true, so in a very clear sense this information about explanation is evidentially relevant. Our thesis does not deny that transparent fact.

As of yet, we have said nothing about what explanation means in proposition (E). Surely some assumptions about this are required by the thesis that explanatoriness is evidentially irrelevant. Here is what we think suffices: even if proposition (E) has entailments about the logical and probabilistic relations of O and H, there is more to explanation than this. Hempel thought that (E) requires that $\Pr(O \mid H)$ is high; in both his deductive-nomological and his inductive-statistical models of explanation, the *explanans* proposition (H) says that the truth of the *explanandum* proposition (O) 'was to be expected'. Salmon (1984) argued against this requirement of high probability, persuasively in our view. But even if you side with Hempel, our argument goes through, provided that you grant the following: for (E) to be true, it isn't sufficient that various logical and probabilistic relations between O and H hold; explanatoriness means something more. We suggest that reasonable candidates for that something more will vindicate our screening-off thesis.

It might be objected that the screening-off test that we have used for confirmational relevance is misguided. Consider proposition I, which says that O logically implies H. Surely I is confirmationally relevant to H. If you know that O is true but don't know that O implies H, and then you learn that this logical fact obtains, that should have an impact on how confident you are that H is true. Yet, O screens off I from H:

$$\Pr(H \mid O) = \Pr(H \mid O\&I).$$

The reason this equality is true is that purely logical and mathematical facts are 'baked into' probability functions. Conditionalizing on them has no impact on posterior probabilities, since (so to speak) they are taken into account from the start. Bayesian confirmation theory standardly assumes that rational agents are 'logically omniscient', an idealization that Garber (1983) tried to set aside in his discussion of the problem of old evidence. We agree that screening-off is not a good test for the confirmational relevance of purely logical facts. But 'if *H* and *O* were true, *H* would explain *O*' is not a purely logical fact, so we stand by our use of the screening-off test in the case at hand. This does not mean that the relevant probability functions can embed only logical truths. Recall that the one used in (4) assumes that proposition *O* is true.

There is a second objection that also alleges that the screening-off criterion is too demanding. Consider a lineage from parent (*p*) to offspring (*o*) to grandoffspring (*g*) where reproduction is uniparental. Suppose that the offspring's having trait *T* screens off the parent's having *T* from the grandoffspring's having *T*:

$$\Pr(g \text{ has } T \mid o \text{ has } T) = \Pr(g \text{ has } T \mid o \text{ has } T \;\&\; p \text{ has } T).$$

It would be wrong to conclude from this that the parent's having trait *T* provides no evidence as to whether the grandoffspring does. It may well. Screening-off does not rule out the possibility that

$$\Pr(g \text{ has } T \mid p \text{ has } T) > \Pr(g \text{ has } T).$$[2]

Our reply is that this point about the lineage is correct, but the problem posed by proposition *E* has a special feature. Recall that we began the paper by considering propositions (1) and (2) as possible representations of the thesis that

[2] Consider a causal chain from ±*X* to ±*Y* to ±*Z*, where each of these is a dichotomous variable. If each state of ±*Y* screens-off each state of ±*X* from each state of ±*Z*, and if +*X* confirms +*Y* and +*Y* confirms +*Z*, then it follows that +*X* confirms +*Z*. See Shogenji 2003. For an equivalent result, see Sober 2009a: 76. For a slightly stronger result, involving a slightly weaker screening-off condition, see Roche 2012.

explanatoriness is confirmationally relevant. We set aside (1) because it is false and (2) because it fails to get at what matters. This led us to the thesis that $\Pr(H \mid O\&E) > \Pr(H \mid O)$ and its contrary (3). It is undeniable that screening-off can be misused as a criterion for confirmational relevance; however, we don't see that it is misused in our focus on the thesis that $\Pr(H \mid O\&E) > \Pr(H \mid O)$ and proposition (3).

We have developed our argument by using the Bayesian theory of confirmation, but we think the lesson generalizes to other theories. Any theory of confirmation that relies just on purely logical and mathematical relations among propositions, and does not use explanatoriness as an evidential principle, will allow a screening-off argument to be developed.

How does our argument concerning the evidential irrelevance of explanatoriness bear on inference to the best explanation (IBE)? IBE is a rule of rational acceptance; it tells you when you should believe that a hypothesis is true (Harman 1965; Lipton 2004; Lycan 2002; Psillos 2007). The version of IBE that we want to discuss holds that explanatoriness is not, in Lycan's apposite phrase, an epistemically irrelevant 'bonbon'; rather, the explanatoriness of a hypothesis is part of what makes it rationally acceptable (in addition to the references just given, see also White 2005).[3] Indeed, it isn't just IBE *theorists* who view IBE in this way; *users* of IBE – for example, in philosophy of science and in metaethics – often take the same view (for examples, see Sober forthcoming).

Acceptance involves a dichotomy – the evidence you have either makes it rational for you to believe the hypothesis or it does not. Whereas IBEists usually are content to think about dichotomous belief, Bayesians usually prefer to think of *degrees* of belief. How are the two concepts related? We suggest the following (standard) connecting principle: if it is rational to believe hypothesis H, based on one's total evidence X, then $\Pr(H \mid X) > 0.5$. There are two factors that determine whether $\Pr(H \mid X) > 0.5$. There is the prior probability $\Pr(H)$ and there is the

[3] Day and Kincaid (1994) and Okasha (2000) have a different understanding of IBE; they consider IBE to be entirely parasitic on a Bayesian calculation of posterior probabilities. Where H_1 and H_2 are competing explanations of one's total evidence X, they say that H_1 is a better explanation than H_2 precisely when H_1 has the higher posterior probability, which Bayes's theorem tells us is true precisely when $\Pr(X \mid H_1)\Pr(H_1) > \Pr(X \mid H_2)\Pr(H_2)$. If this were all that IBE involved, it could not conflict with the Bayesian theory of confirmation, though we then would want to protest that the name of this theory is misleading; a better label would be 'inference to the best hypothesis'.

degree to which *X* confirms *H*. Translated into the language of Bayesianism, this means that if explanatoriness is to influence rational acceptability it must either affect the prior of *H* or the degree to which *X* confirms *H*. We have already argued that explanatoriness should play no role in the Bayesian notion of degree of confirmation. This leaves it open that explanatoriness might be relevant to the assignment of priors; this idea is defended by Lipton (2004), Huemer (2009), and Weisberg (2009). We now will argue that the same reasoning that closes the door for degree of confirmation also slams it shut for prior probabilities.

It is a familiar idea in Bayesianism that today's prior probabilities are often yesterday's posterior probabilities. Our earlier proposition (5) can be used to illustrate this relation of present to past. If you learn today that *S* got lung cancer after age 50 (and this is all you learn), your new prior has the same value as your old posterior:

$\mathrm{Pr}_{\mathrm{today}}$(*S* smoked at least 10,000 cigarettes before age 50) = $\mathrm{Pr}_{\mathrm{yesterday}}$(*S* smoked at least 10,000 cigarettes before age 50 $\vert$ *S* got lung cancer after age 50).

In connection with the second of these probabilities, we argued before that facts about explanatoriness get screened-off; adding a claim about explanation to the conditioning proposition does not change the value of yesterday's probability. This means that the value of today's prior is also unaffected by explanatoriness.

We said above that today's priors are 'often' yesterday's posteriors. We said 'often' to set *first priors* to one side; perhaps there are some (nontrivial) prior probabilities that are rock bottom, not based on any observational evidence at all. Should first priors be assigned values to reflect considerations of explanatoriness? We are sceptical. The explanatoriness of a hypothesis has to do with its relationship to other propositions. For example, the hypothesis (*H*) that *S* was a heavy smoker will be very explanatory if *S* gets lung cancer, but *H* will be much less explanatory if *S* fails to contract lung cancer. If you don't know whether *S* contracts lung cancer, you don't know whether *H* is very explanatory or very unexplanatory. In this state of ignorance, should you assign *H* a high prior because it will be very explanatory if *S* gets lung cancer, or should you assign *H* a low prior because it will be very unexplanatory if *S* fails to get lung cancer?[4] First priors are supposed to be assigned on the basis of zero observational information. We are okay with tautologies and contradictions being assigned priors in this circumstance (though this has nothing to do with explanatoriness). But how this meagre basis can be augmented by bringing in ‘explanatoriness’ is a mystery to us.[5]

Our screening-off thesis is related to Van Fraassen’s (1989) thesis that IBE is probabilistically incoherent and therefore subject to a Dutch book. Van Fraassen thinks that IBE proposes a two-step rule for updating: if the evidence *O* increases *H*’s probability, then *H* receives a further boost in probability if *H* would provide a good explanation of *O*. Our argument aims to show that the explanatoriness of *H* cannot provide this additional boost; in addition, it sidesteps the question of how the apparently prudential considerations introduced by Dutch book arguments are relevant to a non-prudential notion of rational degree of belief.

Friends of inference to the best explanation may be tempted to draw the following conclusion from our argument: *so much the worse for Bayesianism*. IBEists who reject Bayesianism because they think that explanatoriness is confirmationally relevant need to formulate a nonBayesian theory of confirmation. Even if they think that explanatoriness is confirmatory, this supplies only a sufficient condition for confirmation; it is perfectly clear that *O* can confirm *H* even when *H* if true would not explain *O*. Once necessary and sufficient conditions for confirmation are specified, the challenge is to show why explanatoriness is evidentially relevant.[6]

Texas Christian University
Fort Worth, TX 76129, USA
w.roche@tcu.edu

[4] As mentioned earlier, Salmon (1984) rejects the thesis that the *explanans* must show that the *explanandum* was to be expected. This does not help IBEists who want the assignment of values to first priors to be influenced by considerations of explanatoriness.

[5] Can simplicity be used to justify an assignment of first priors? Sober (2009b) argues that scientifically legitimate uses of parsimony and simplicity rest on empirical assumptions.

[6] We thank Branden Fitelson, Mike Titelbaum, and Danielle Wylie for useful discussion.

University of Wisconsin, Madison

Madison, WI 53706, USA

ersober@wisc.edu

References

Bromberger, S. 1966. Why questions. In *Mind and Cosmos*, ed. R. G. Colodny, 86–111. Pittsburgh, PA: University of Pittsburgh Press.

Day, T. and H. Kincaid. 1994. Putting inference to the best explanation in its place. *Synthese* 98: 271–95.

Fisher, R. A. 1959. *Smoking: the Cancer Controversy – Some Attempts to Assess the Evidence*. Edinburgh, Scotland: Oliver and Boyd.

Fitelson, B. 1999. The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science* 66: S362–78.

Garber, D. 1983. Old evidence and logical omniscience in Bayesian confirmation theory. In *Minnesota Studies in the Philosophy of Science (vol. 10): Testing Scientific Theories*, ed. J. Earman, 99–131. Minneapolis: University of Minnesota Press.

Harman, G. 1965. The inference to the best explanation. *Philosophical Review* 74: 88–95.

Hempel, C. G. 1965. Aspects of scientific explanation. In his *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 331–496. New York: Free Press.

Huemer, M. 2009. Explanationist aid for the theory of inductive logic. *British Journal for the Philosophy of Science* 60: 345–75.

Lipton, P. 2004. *Inference to the Best Explanation*, 2nd edn. London: Routledge.

Lycan, W. 2002. Explanation and epistemology. In *Oxford Handbook of Epistemology*, ed. P. Moser, 408–33. Oxford: Oxford University Press.

Okasha, S. 2000. Van Fraassen's critique of inference to the best explanation. *Studies in the History and. Philosophy of Science* 31: 691–710.

Psillos, S. 2007. The fine structure of inference to the best explanation. *Philosophy and Phenomenological Research* 74: 441–48.

Roche, W. 2012. A weaker condition for transitivity in probabilistic support. *European Journal for Philosophy of Science* 2: 111–18.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Shogenji, T. 2003. A condition for transitivity in probabilistic support. *British Journal for the Philosophy of Science* 54: 613–16.

Sober, E. 2009a. Absence of evidence and evidence of absence – evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies* 143: 63–90.

Sober, E. 2009b. Parsimony arguments in science and philosophy – a test case for naturalism. *Proceedings and Addresses of the American Philosophical Association* 83: 117–55.

Sober, E. forthcoming. Two Cornell realisms – moral and scientific. *Philosophical Studies*.

Van Fraassen, B. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

Weisberg, J. 2009. Locating IBE in the Bayesian framework. *Synthese* 167: 125–43.

White, R. 2005. Explanation as a guide to induction. *Philosophers' Imprint* 5.