Colin Howson

How pseudo-hypotheses defeat a non-Bayesian theory of evidence: reply to Bandyopadhyay, Taper, and Brittan

Article (Accepted version) (Refereed)

Original citation:
Howson, Colin (2016) *How pseudo-hypotheses defeat a non-Bayesian theory of evidence: reply to Bandyopadhyay, Taper, and Brittan.* International Studies in Philosophy of Science, 30 (3). pp. 299-306. ISSN 0269-8595

DOI: 10.1080/02698595.2017.1316111

DISCUSSION NOTE (forthcoming in International Studies in the Philosophy of Science)

Reply to Bandyopadhyay, Taper, and Brittan

Colin Howson
Department of Philosophy, Logic and Scientific Method, London School of Economics

ABSTRACT
Bandyopadhyay, Taper, and Brittan (BTB) advance a measure of evidential support that first appeared in the statistical and philosophical literature four decades ago and has been extensively discussed since. I have argued elsewhere, however, that it is vulnerable to a simple counterexample. BTB claim that the counterexample is flawed because it conflates evidence with confirmation. In this reply, I argue that the counterexample stands, and is fatal to the doctrine of likelihoodism.

1. Introduction

A focus of interest in formal philosophy of science for well over half a century has been the analysis of the idea of the *amount of evidential support for a hypothesis*, or the *evidential strength in favour of a hypothesis*, which some given set of observational data affords. Just what form an admissible measure should take, however, has been a subject of controversy, with no universally accepted candidate emerging. According to one widespread view, the Bayesian view, it is inseparable from considerations of changes in belief induced by the data, where belief is of course measured by a probability function (and hence from considerations of confirmation: the Bayesian qualitative criterion of confirmation is that $P(H|E) > P(H)$). Thus Jan Sprenger: 'Evidential support is based on comparing past and present degrees of belief' (Sprenger 2014, 7). This does however raise the question of how the comparison should be rendered in terms of a measure: should that be a simple difference, or a ratio, or some other function? In a well-known article, Ellery Eells and Branden Fitelson (2002) evaluated several candidate measures against a list of what they took to be relevant criteria. One of those measures, scoring equal highest on Eells and Fitelson's criteria and which will figure strongly in the subsequent discussion, was the logarithm of the likelihood-ratio (henceforward LR), $P(E|H) / P(E|{\sim}H)$. Eells and Fitelson's article sparked a strong revival of interest in measures of evidential support, and a good deal of further work in the area was published, including the very systematic investigation by Vincenzo Crupi, Katya Tentori, and Michel Gonzalez (2007), a novel feature of which is an empirical comparison contributing, as the authors see it, to the psychology of inductive reasoning.

The LR is distinctive in being simply a coefficient of proportionality relating prior and posterior beliefs in the odds form of Bayes's theorem:

$$\mathrm{Odds}(H|E) = \mathrm{LR} \times \mathrm{Odds}(H) \qquad (1)$$

Logging both sides gives the additive form

$$\mathrm{lodds}(H|E) = \mathrm{logLR} + \mathrm{lodds}(H)$$

where 'lodds' is short for 'logOdds'. The logarithm with a base greater than 1 is an increasing function, so logging the LR gives an ordinally equivalent measure that adds over conditionally independent pieces of evidence.[1] Since $\mathrm{Odds}(H|E) > \mathrm{Odds}(H)$ if and only if $P(H|E) > P(H)$, equation (1) tells us immediately that a necessary and sufficient condition for H to confirm E Bayesianwise is that $\mathrm{LR} > 1$, assuming of course that $\mathrm{Odds}(H)$, and hence $P(H)$, is nonzero.

Probably the first to note the relationship displayed in (1) between the LR and prior and posterior odds was the eminent Bayesian pioneer, geophysicist, and statistician, Sir Harold Jeffreys. He called logLR the 'support' for H by E, but he attached a caution to that use of the measure: 'It is reasonably clear that [H] has a moderate prior probability in practical cases, for if it had one near 0 we should not consider its truth worth investigating' (Jeffreys 1936, 416; in 'practical cases' he took the priors of H and ${\sim}H$ as 0.5). In the light of the challenge of Prasanta S. Bandyopadhyay, Mark L. Taper, and Gordon Brittan, Jr. (2016; henceforward BTB) to me, we shall see that the caution was prescient.

I. J. Good later called the LR the 'weight of evidence' in favour of H (Good 1983, 36; following, as he tells us, the lead of Alan Turing, Good thought 10 the most appropriate base for the logarithm). ${\sim}H$ is of course just one alternative hypothesis to H, so it is natural to generalise the measure to an arbitrary pair H, H' (where H' does not even need to be inconsistent with H) as $P(E|H) / P(E|H')$, or its log as did Good (1983, 159).[2] In this case we can rewrite (1) as

$$\text{Odds}(H, H'|E) = \text{LR} \times \text{Odds}(H, H') \quad (2)$$

where Odds(H, H') = P(H) / P(H'); similarly for the conditional case.

2. Likelihoodism

There is another camp, comprising people often called *likelihoodists*, which also adopts the LR as a measure of evidential support, but repudiates any justification of it in terms of incremental belief: on the contrary, likelihoodists see the LR as an objective, stand-alone measure of evidential support independent of any beliefs anyone might have in the hypotheses so compared. This view, which had its roots in the writings of R. A. Fisher, became widely publicised by the geneticist and statistician Anthony Edwards in an influential book, in which he proclaimed it 'The Likelihood Axiom' (Edwards 1972). He was followed by, among others, the statistician Richard Royall (1997) and the philosopher of science Elliott Sober (2002). For later reference, here is Edwards's statement of his axiom:

> Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other. (Edwards 1972, 31; emphasis in the original)

It can be quickly verified from an inspection of BTB's text that Edwards's axiom is identical, apart from details of wording, to what BTB describe, without mentioning Edwards, as 'our account of evidence based on the likelihood-ratio measure' (BTB 2016, 3; 'their' account is that the LR is an objective, belief-independent measure of the relative strength of evidence for pairs of statistical hypotheses—i.e. the Likelihood Axiom). Nor do BTB cite any of the very extensive literature comparing the merits of functions of the evidence other than the LR as measures of evidential support. Instead they offer their own justification of the LR:

> The first reason for choosing LR is that it is the most *efficient* evidential function in the sense that we can gather strong evidence for the best model with the smallest amount of data. Second, LR as a measure is insensitive to the choice of priors, whereas the posterior and prior difference measure and

> many other measures of confirmation are not (see section 4 for an example). This captures the idea of evidence being objective, i.e. not unduly influenced by the agent's prior probability. Third, compared to some other measures of evidence [they cite Christensen and Joyce] ... LR, as a measure of evidence, does not behave erratically ... If any measure of evidence fails to reflect the change in the probability of data when the priors of the hypotheses do not have any effect on any such measure, then the latter can be branded as an inadequate measure of evidence. In contrast, LR does not have such a counterintuitive consequence. And fourth, likelihoods and the LR are already embedded in the Bayesian approach and this familiarity makes the distinctions easier to grasp. (BTB 2016, 4–5)

None of these reasons is compelling, and three can be immediately discounted. Consider the fourth: for authors who propose the LR as a non-Bayesian measure of evidential support, to cite its embedding in the Bayesian approach as a point in its favour verges on the bizarre. The first reason can also be ignored, begging as it does the question that the LR is a measure of evidential strength at all. The second reason actually *rules out* the LR measure when alternatives of the form H and $\sim H$ are permitted into its domain. At the beginning of their article, BTB say that they will restrict their account to so-called simple statistical hypotheses, i.e. statistical hypotheses with all parameter values specified and hence conferring a definite probability on the data, but the restriction is later relaxed with one of the pairs allowed to be the negation of the other (BTB 2016, 4).[3] But admitting LRs of the form $P(E|H) / P(E|\sim H)$ means a dependence on priors, since the probability calculus tells us that $P(E|\sim H) = P(\sim H)^{-1} \Sigma_i P(E|H_i)P(H_i)$ for every finite partition of $\sim H$. In the light of this observation, the third reason in the list above is simply not applicable.

BTB and likelihoodists generally may want to divorce their measure from any consideration of prior probabilities, but given the preceding paragraph this seems unfeasible without crippling restrictions on its domain. BTB also claim that 'very strong evidence [as equated by BTB with a large LR] for a hypothesis does not entail that it is more believable' (BTB 2016, 3), but inspection of (1) undermines that assertion. For suppose LR is large, in the minimal sense of exceeding unity. It immediately follows from (1) that $Odds(H|E) > Odds(H)$ which entails that $P(H|E) > P(H)$. Similarly, we can quickly infer from (2) that

$$P(E|H) / P(E|H') = [P(H|E) / P(H)] \div [P(H'|E) / P(H')].$$

So a LR greater than 1 implies that the ratio increment of belief in *H* exceeds the ratio increment of belief in *H*', whence it follows that if the priors of *H* and *H*' are equal the LR is just the ratio of *posterior probabilities.* These elementary demonstrations assume positive priors, but that is something that BTB (2016, 4) assume.

I think it does no good to claim, as does Sober (2002, 25), that likelihoodists can ignore these facts because they are not committed to a probabilistic, i.e. Bayesian, theory of partial belief. Pioneering work by Bruno de Finetti and L. J. Savage famously showed that if an agent evaluates courses of action under uncertainty according to intuitively compelling criteria of consistency, they will behave as if they were expected utility maximisers equipped with a prior probability distribution over a state-space. Indeed, it is on this foundation that modern Bayesian theory squarely rests. One may criticise the consistency postulates employed by de Finetti and Savage, but they have been successively refined since those classic papers were published, and the claim that consistent beliefs should obey at least the finite probability axioms is now so widely accepted that to say you simply don't accept that view is not an adequate response.

Armed with these preliminary observations, we can now turn to the counterexample whose cogency BTB's article challenges.

3. The Counterexample

I first advanced the example against Edwards's Likelihood Axiom in Howson (2002), except that there 'God' replaced 'Santa', and later against Stathis Psillos's similar claim that LRs *by themselves* 'capture the comparative impact of the evidence on competing hypotheses' (Psillos 2009, 67; Howson 2013). Here is the counterexample, quoted by BTB from my 2013 discussion:

> Consider a hypothesis *H* that everyone agrees is plainly false, for example, that Santa willed the outcome of 100 flips of a fair-looking coin. If the outcome is specified in *H* then [the reciprocal of the likelihood ratio], where ~*H* is plausibly the hypothesis that the coin is approximately fair, is very small indeed ($2^{100} = 1.267 \times 10^{30}$), but that is not sufficient to induce any feeling that *H* might be true. (Howson 2013, 209)

It should be clear that the example is intended to show that any serious explanatory theory can always be weakly or strongly dominated in terms of the LR by any preposterous pseudo-hypothesis cooked up to deliver the data, and that therefore the LR cannot sensibly be interpreted *by itself* as a measure of relative evidential strength. But BTB's paraphrase of the quotation completely misstates my position:

> The crux of Howson's argument against a non-subjective account of evidence is that although everyone believes that the Santa hypothesis is false, i.e., highly implausible, our account of evidence based on the likelihood-ratio measure provides very strong evidence for it. He is right to point out that in his example the evidential support for *H* against ~*H* equals 1.267 x 10^{30}, which signifies very strong evidence for the Santa hypothesis against the fair-coin hypothesis. So, he concludes, that there must be something wrong with this account, on which very strong evidence for a hypothesis does not entail that it is more believable. (BTB 2016, 3)

I clearly did *not* 'point out' that the evidential support for *H* against ~*H* is equal to 1.267 x 10^{30}, nor does the quotation from me in any way justify the claim that I believe the data are very strong evidence for *H*. On the contrary, as I remarked above, the point of the example is to illustrate that because of its very nature the data provide no genuine evidence *whatever* for *H*, a view of the matter I underlined by saying that *H* was 'plainly false'. 'Plainly false' here is most plausibly construed in terms of a prior probability whose value is *effectively* zero, in the sense of being actually zero or positive but infinitesimal.[4] It follows immediately from (1) that any increase in such a prior probability induced by the LR, *however large* that might be, is effectively zero.

For a Bayesian this disposes of the challenge of facile 'hypotheses' like Santa, whose credibility is zero and whose fit to the data is guaranteed in advance. *But for the likelihoodist like BTB there is no way out*. An obvious consequence of their position is that it is trivially easy to outscore any seriously proposed explanatory theory. Indeed, any sufficient condition for the data would *strictly* dominate any theory giving the data less than 100% probability, which of course includes any purely statistical hypothesis. Interpreting LRs as stand-alone measures of relative evidential strength therefore not only flies in the face of common sense[5] but makes a nonsense of scientific methodology. BTB may claim that such counterexamples 'conflate confirmation with evidence', but what else is evidence all about? Other things being equal, if the evidence is stronger for one theory than for another the former is more believable, given that evidence. This is a judgement the Bayesian theory corroborates in its result, noted earlier, that if the priors are equal and nonzero so are the posterior probabilities. On the other hand, assigning an effectively zero prior probability precludes any such absurd judgement as that {'The moon is made of cheese'}$\cup${*M*} gets comparable evidential support with Einstein's field equations from the Mercury perihelion observations (*M*).[6]

Sober tried to pre-empt this way of dismissing gerrymandered pseudo-explanations by claiming that Bayesians will assign them only an intermediate prior probability because 'Bayesians usually reserve priors of 0 or 1 for tautologies and contradictions' (Sober 2002, 25n6). Even were that statement true, it ignores the possibility of assigning infinitesimal values. But it is not true: though some Bayesians (notably the late Dennis Lindley) have argued in favour of restricting the extreme probability values in the way Sober describes, that view[7] has never, for what I believe are very good reasons,[8] been part of 'official' Bayesian doctrine. As the Bayesian Wesley C. Salmon pointed out, 'we have good reasons for avoiding the assignment of extreme values to the priors of the hypotheses *with which we are seriously concerned*' (Salmon 1991, 184; emphasis added); and Santa is of course hardly a hypothesis with which we are seriously concerned. Recall also Jeffreys's caution that the prior probabilities in the LR be not too small when the latter is interpreted as strength of evidence. BTB take a similar line to Sober's (without however mentioning him), claiming that Bayesians do not assign the extreme values to what BTB call 'empirical hypotheses' because doing so would prevent learning from experience via the rule of conditionalisation (BTB 2016, 4).[9] But I doubt that any Bayesian would want to 'learn' Santa by conditionalising on the data that Santa was constructed to entail. Santa's not for learning, as one might put it.

But BTB have not yet finished with their attempt another to undermine the Santa counterexample:

> On our account of evidence, we always compare the merits of two hypotheses.[10] For a different pair of hypotheses, e.g., the coin being double-headed and the Santa hypothesis, given the outcomes of 100 heads out of 100 flips, both hypotheses are equally supported. We cannot distinguish the double-headed hypothesis from the Santa hypothesis given the data. Therefore, it does *not* follow that given the data we will *always* find very strong evidence for the Santa hypothesis as it depends on what two competing hypotheses we are dealing with. (BTB 2016, 7; emphases in the original)

Note the sadly characteristic disregard of my own words: I certainly did not claim that 'we would always find strong evidence for the Santa hypothesis', and BTB's emphasising 'not' and 'always' makes their misreporting even less pardonable. Distortion apart, this quotation reveals a strange misunderstanding of what a counterexample actually is. If one pair of hypotheses is a counterexample to a claim, the possibility that a different pair can be found that is not is, of course, quite irrelevant.

Not content with trying to demolish Santa, BTB mount a challenge to another part of the same paper of mine from which that example was drawn, where I wrote, 'If anything characterises Bayesians ... it is the conviction that no satisfactory reconstruction of scientific inference is possible *without* appealing to prior probabilities' (Howson 2013, 208; emphasis in the original). Not only do BTB misquote this into the *inequivalent* sentence, 'it is the conviction [of the author] that no satisfactory reconstruction of scientific inference is possible without appealing to prior probabilities' (also dropping my emphasis), but they then completely misrepresent it: citing the fact that Pauli and others were prepared to bet on the truth of parity conservation prior to the experiments that disproved it, BTB say, 'This episode is evidence that the invocation of subjective prior probabilities can sometimes be detrimental to scientific progress, contrary to what Howson claims (see his sentence quoted in section 2 for it)' (BTB 2016, 8). The 'sentence quoted in section 2' is the sentence misquoted above, which is clearly *not* a claim that invoking prior probabilities is never detrimental to scientific progress: my claim was that no *reconstruction of theory evaluation* in the sciences can ignore prior probabilities.

4. Conclusion

BTB end their paper with this claim:

> We launched a three-pronged argument against Howson. First, we demonstrated that his counterexample against our likelihood ratio–based account of evidence conflates confirmation with evidence. Second, for a different pair of hypotheses, we showed his counterexample does not work. Third, in scientific practice he ignores the role of evidence, which, we argued, is agent independent. (BTB 2016, 8)

I hope that in the light of the foregoing the reader can accept that none of these charges comes anywhere near being substantiated. The third, which on the face of it is simply preposterous, in fact garbles a claim BTB made more clearly earlier in their paper, which is that I

> ignore a fundamental part of scientific inference in which data's providing evidence for one hypothesis against its alternative does not entail a belief that it is true. (BTB 2016, 8)

But since I never said that providing evidence for one hypothesis against its alternative does entail a belief that it is true, this charge is one more misrepresentation among a depressing multitude.

I conclude on a more general note. Although likelihoodism has seemed to some a plausible alternative to classical frequentism (hypothesis testing *à la* Fisher or Neyman-Pearson, plus associated estimation criteria) and Bayesianism, unlike either of those two it is fatally vulnerable to the easy manufacture of pseudo-hypotheses like the Santa example, gerrymandered to fit the data but (of course) meriting no credit for so doing.

References

Bandyopadhyay, P., M. Taper, and G. Brittan, Jr. 2016. "Non-Bayesian Accounts of Evidence: Howson's Counterexample Countered." *International Studies in the Philosophy of Science* 30: xxx–xxx.

Bernstein, A. R., and F. Wattenberg. 1969. "Nonstandard Measure Theory." In *Applications of Model Theory to Algebra, Analysis, and Probability*, edited by W. A. J. Luxemburg, 171–186. New York: Holt, Rinehart and Winston.

Crupi, V., K. Tentori, and M. Gonzalez. 2007. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74: 229–252.

Edwards, A. W. F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Eells, E., and B. Fitelson. 2001. "Symmetries and Asymmetries in Evidential Support." *Philosophical Studies* 107: 129–142.

Good, I. J. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Howson, C. 2002. "Bayesianism in Statistics." In *Bayes's Theorem*, edited by R. Swinburne, 39–70. Oxford: Oxford University Press.

Howson, C. 2013. "Exhuming the No-miracles Argument." *Analysis* 73: 205–211.

Howson, C. 2017. "Repelling a Prussian Charge with the Solution to a Paradox of Dubins." *Synthese*, online September 2016: 1-9.

Jeffreys, H. 1936. "Further Significance Tests." *Proceedings of the Cambridge Philosophical Society* 32: 416–445.

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

Psillos, S. 2009. *Knowing the Structure of Nature: Essays on Realism and Explanation*. Houndmills: Palgrave Macmillan.

Royall, R. 1997. *Statistical Evidence—A Likelihood Paradigm*. London: Chapman and Hall.

Salmon, W. C. 1990. "Rationality and Objectivity in Science, or Tom Kuhn meets Tom Bayes." In *Scientific Theories*, edited by C. W. Savage, 175–204. Minneapolis: University of Minnesota Press.

Sober, E. 2002. "Bayesianism—Its Scope and Limits." In *Bayes's Theorem*, edited by R. Swinburne, 21–38. Oxford: Oxford University Press.

Sprenger, J. 2014. "From Evidential Support to a Measure of Corroboration." PhilSci-Archive. http://philsci-archive.pitt.edu/10753.

Notes

[1] Another measure ordinally equivalent to the LR is $[P(E|H) - P(E|{\sim}H)] \div [P(E|H) + P(E|{\sim}H)]$. This function was derived from their own list of adequacy criteria for measures of factual support by Kemeny and Oppenheim (1952).

[2] Good also explicitly included a symbol K to signify background information. Nowadays most authors regard this information as absorbed in P.

[3] Allowing only simple hypotheses would preclude the enormously wide range of applications where simple hypotheses are tested against composite alternatives: e.g. a simple null hypothesis specifying a parameter value $t = t_0$ against a composite alternative like $t \neq t_0$ or $t > t_0$ or $t < t_0$. But why should a measure of evidential support be restricted to statistical hypotheses, and not theories generally? BTB do not say. The Santa hypothesis at the centre of my counterexample is not statistical but deterministic, so it may be that BTB mean to include such hypotheses as limiting cases (as we shall see, Santa predicts the data with probability 1).

[4] An infinitesimal is a number less in absolute magnitude than any positive real number. I should emphasise that these are perfectly respectable numbers. In the mid-twentieth century, Abraham Robinson exploited the resources of model theory to show that there is an elementary extension of the real numbers, a field of so-called *hyperreal numbers*, containing besides copies of the real numbers themselves, infinitesimals and infinite reciprocals of infinitesimals. *Nonstandard analysis*, *nonstandard probability theory and nonstandard physics* have since become flourishing fields in their own right, offering often much simpler proofs of classical results. A well-known result of Bernstein and Wattenberg (1969) is that positive infinitesimal probability values can be assigned to all the members of a continuum-sized outcome-space in such a way that the values sum, in a suitable sense, to 1.

[5] We can note that the great Bayesian, Laplace, famously described the Bayesian theory as common sense reduced to a calculus—'le bon sens réduit au calcul'.

[6] Ironically BTB themselves find the conflation they condemn irresistible: a positive outcome from a diagnostic test for TB in which the LR is 25.7 shows, they say, 'that the individual is more likely (*approximately 26 times more likely*) to have the disease than not' (BTB 2016, 6; emphasis added).

[7] It is known as the principle of Regularity, so named by Rudolf Carnap. The technical problem for Regularists is that many of the outcome spaces of statistics are intervals of real numbers, and it is mathematically impossible to assign positive real-valued

probabilities to all the points in them. There have been appeals to infinitesimals to solve the problem (see note 5 above), though these attempts are controversial. Howson (2017) is my own contribution to the discussion.

[8] Mentioned in Howson (2017).

[9] It is easy to show that no evidence can increase a probability of 0, or for that matter 1, by that means. BTB also misstate the principle, which they claim 'says that [the agent's] degree of belief in H_1 after the data are known is given by the conditionalisation principle $\Pr(H_1|D)$, assuming the $\Pr(D)$ is not zero' (BTB 2016, 4). But $\Pr(H_1|D)$ is a *number*, not a principle. What the principle does say is that after learning D, and nothing stronger, the agent's new belief function should be $P_D(\ .\ ) := P(\ .\ |D)$. Nor is there any need to assume that $P(D)$ is nonzero: there are well-known axiomatisations of conditional probability in which the second argument can have probability 0, so long as it is consistent.

[10] Recall that this provision is part of Edwards's Likelihood Axiom!