The Bayesian Treatment of Auxiliary Hypotheses: Reply to Fitelson and Waterman

Michael Strevens

To appear in *British Journal for the Philosophy of Science*

Abstract

Fitelson and Waterman (2004)'s principal objection to Strevens (2001)'s Bayesian treatment of auxiliary hypotheses rests on a misinterpretation of Strevens's central claim about the negligibility of certain small probabilities. The present paper clarifies and proves a very general version of the claim.

1. The Project

Fitelson and Waterman (2004) argue that the Bayesian treatment of auxiliary hypotheses proposed in Strevens (2001) is inadequate; in what follows, I offer a defense.

Let me begin by summarizing the argument in my original paper, as Fitelson and Waterman's overview is potentially misleading in several important respects. I attempt to avert any misinterpretation in section 2, and then address in section 3 what is by far the most important part of Fitelson and Waterman's critique, showing that their theorem 2 does not in any way undercut the main claim of my treatment of auxiliary hypotheses.

Suppose that a main hypothesis *h* and an auxiliary hypothesis *a* together assign a physical probability to a piece of evidence *e*, just as envisaged in the usual Quine-Duhem problem. What factors determine the way in which the observation of *e* impacts the probabilities of *h* and *a*?

There can be no general Bayesian treatment of this question, because the Bayesian apparatus allows that, in different circumstances, *e* impacts on *h* and *a* in just about any way you like, essentially because *e* may bear not only on *ha* as a "corporate body" (to use Quine's expression) but also separately on *h* and *a*. A fruitful Bayesian approach must restrict itself to some particular class of cases that is interesting to confirmation theorists and in which there is some systematic pattern in *e*'s impact on *h* and *a*—if there is such a class.

In Strevens (2001), I take the following route to a satisfying treatment. Divide the effect of *e* on *h* and *a* into two parts: first, the effect due to *e*'s impact on *ha* as a corporate body, and second, the remainder—the effect presumably due to *e*'s additional impact on *h* and *a* separately. Explore the properties of the first kind of impact in isolation. (I argue that discovering these properties is, in fact, the real Quine-Duhem problem, but I put that issue aside here.) They turn out to be systematic in an interesting and suggestive way (sections 3 and 4 of my original paper). Then explore the conditions under which the second kind of impact is negligible, so that the systematic properties of the first kind of impact dominate the total effect of conditionalizing on *e* (section 5 of my original paper).

I show that the second kind of impact is negligible in a class of cases very important to the study of confirmation in science, namely, those where the auxiliary hypothesis concerns the experimental conditions under which *e* is produced. (I refer the reader to the original paper for the details.) In these cases, then, something interesting and substantive can be said about the impact of *e* on *h* and *a*. What can be said turns out to include the following:

1. When *e* falsifies *ha*, the "blame" is distributed between *h* and *a* roughly

in proportion to their relative prior probabilities, so that a more probable h will be blamed relatively less.

2. The magnitude, positive or negative, of the impact of e on the main hypothesis h is greater the more probable the auxiliary a. When the probability of a is high, favorable evidence provides a greater boost to the probability of h, whereas unfavorable evidence makes a bigger dent in h's probability.

These results are then applied to the problem of "ad hoc" auxiliary hypotheses.

The main point of Fitelson and Waterman's reply is that what I will call the *negligibility argument*—my argument for the negligibility of the second kind of impact—is flawed. I defend the argument in section 3.

2. Clarifications

There are a number of ways in which Fitelson and Waterman's presentation of my view might mislead the reader. First, they say (§2) that I assume the logical equivalence of e and $\neg(ha)$. In fact, I need nothing anywhere near that strong; what I demonstrate (not merely assume) is a kind of local "probabilistic equivalence", local because it concerns only the effect of e on h and a. Thus my treatment involves no "logical weakening" of the evidence, contrary to Fitelson and Waterman's claim, and the statement of equivalence—of probabilistic equivalence—far from being an "idealization", is true, or approximately so, for the class of cases to which my approach is intended to apply.

Second, Fitelson and Waterman claim (§4) that my treatment is intended to cover "all interesting Quine-Duhem cases". Not so; the treatment is explicitly restricted to the large class of cases briefly characterized in the previous section and described at greater length in section 5 of the original paper.

These are the cases for which the probabilistic equivalence relation is shown to hold.

Third, a reader of Fitelson and Waterman's main text alone would think that the negligibility argument is confined to cases in which *e* falsifies *ha*. It is far more general than this; it applies to cases in which *e* has any probabilistic impact whatsoever on *ha*. (Fitelson and Waterman point out in their footnote 1 that it applies to any negative impact; it applies to any positive impact as well.)

Fourth, by restricting their attention to the question of whether *h* or *a* is relatively more confirmed or disconfirmed by *e*, Fitelson and Waterman ignore the most interesting claims in the paper, such as claim (2) from the previous section.

Fifth, Fitelson and Waterman and I differ considerably in our interpretation of the Quine-Duhem problem. I hold that the problem is principally concerned with evidence that impacts on *h* and *a* only by impacting on *ha*. Fitelson and Waterman appear to think otherwise (see the comments following their Theorem 1). This perhaps explains a number of our disagreements; however, for the purposes of this reply, I put aside the question of the proper interpretation of the Quine-Duhem problem altogether, and consider my original paper as answering the question above: what are the interesting properties of cases in which the hypothesis to be tested *h* needs to be supplemented with an auxiliary hypothesis *a* in order to determine a physical probability for the evidence *e*?

Sixth, Fitelson and Waterman suggest (§3) that I use a ratio measure of degree of confirmation and then later (§4) that I regard posterior probability as a good measure of degree of confirmation. It is true that I discuss both ratios of posteriors and, a fortiori, posteriors themselves. But at no point do I endorse either as a measure of degree of confirmation.

Like most Bayesians, but unlike Fitelson and Waterman, I do not consider the selection of a single correct measure of confirmational relevance

essential for work in confirmation theory. All locutions in my paper that suggest to Fitelson and Waterman the use of such a measure (e.g., "e impacts more negatively on h than on a") should be regarded as verbal paraphrases of mathematical facts about the dynamics of probability under conditionalization. It is for this reason that my preferred medium of comparison is, where possible, the graph (figures 1 and 2 of my original paper).

3. The Negligibility Argument

Fitelson and Waterman take issue with my argument that, for an important class of cases, the impact of e on h and a is almost entirely contained in its impact on the corporate body ha. They focus on the special case in which ha entails $\neg e$, that is, in which the evidence e falsifies ha.

The negligibility argument depends on a certain claim, which in Fitelson and Waterman's special case amounts to the following:

> If the two prior probabilities $P(e|h\neg a)$ and $P(e|\neg(ha))$ are close, then the posterior probability $P^+(h)$ after conditionalizing on e is approximately equal to $P(h|\neg(ha))$.

(Fitelson and Waterman formulate everything as priors, and so write $P(h|e)$ where I have $P^+(h)$. The posterior notation makes the mathematics tidier and, I think, a little easier to follow.)

Fitelson and Waterman's theorem 2 shows that the claim is false, when interpreted in a certain way. But the claim is not interpreted this way in my original paper; on the original interpretation, it is demonstrably true, as I now show.

Fitelson and Waterman write the all-important claim as follows (as above, I substitute a posterior probability for their prior probability, simply a notational variant): If $P(e|h\neg a) \approx P(e|\neg(ha))$ then $P^+(h) \approx P(h|\neg(ha))$. This is incorrect, as Fitelson and Waterman show, if, as their chosen notation

suggests, the two approximate equality relations are given the same interpretation, and in particular, if they are interpreted as approximate equalities of difference, so that two quantities x and y are approximately equal just in case $x - y = \epsilon$ for some small ϵ, positive or negative.

The claim holds, however, if the first approximate equality is interpreted as one of ratio, so that two quantities x and y are approximately equal just in case x/y is close to one, or more exactly, just in case $x/y = 1 + \epsilon$ for some small ϵ, positive or negative, and the second approximate equality is interpreted as one of difference.

More formally, the following result can be proved: for any ϵ, positive or negative, if

$$\frac{P(e|h\neg a)}{P(e|\neg(ha))} = 1 + \epsilon$$

then

$$P^{+}(h) - P(h|\neg(ha)) \leq \epsilon.$$

This is the basis of my negligibility argument in the special case considered by Fitelson and Waterman.

A stronger theorem, proved in the next section, shows that the same approximation claim holds for the general case where e impacts on ha probabilistically.

I regret not making the nature of the approximate equality clearer in the original article (though the importance of measuring approximate equality by taking the ratio was stated explicitly on p. 532).

One further issue ought to be mentioned. When the probabilities mentioned in the negligibility argument are very small, they may be approximately equal in the difference sense but not in the ratio sense. Extra care must therefore be taken in applying the argument in such cases. The implications for my approach to the Quine-Duhem problem are discussed at length in the original paper (see pp. 532–533).

4. Generalization and Proof

Theorem *For any real-valued ϵ, if*

$$\frac{P(e|h\neg a)}{P(e|\neg(ha))} = 1 + \epsilon$$

then

$$P^+(h) - Q(h) \leq \epsilon$$

where $Q(h)$ is defined as $P^+(ha) + P(h|\neg(ha))P^+(\neg(ha))$ and $P^+(\cdot) = P(\cdot|e)$.

Note that there is no constraint on ϵ. In Fitelson and Waterman's special case where ha entails $\neg e$, $Q(h) = P(h|\neg(ha))$, so this theorem subsumes the result stated in the previous section.

When the conditions stated by the theorem obtain, then, $P^+(h)$ will behave like $Q(h)$. In particular, since $Q(h)$ exhibits the behavior described in the numbered clauses in section 1—increasing by more when $P(a)$ is high, if e boosts the probability of ha, for example—$P^+(h)$ will do so too, within the margin of error allowed by ϵ. (The behavior of $Q(h)$ is investigated in sections 3 and 4 of my original paper.)

Observe that this claim does not assume any particular measure of degree of confirmation. Nor does it depend on any particular interpretation of $Q(h)$, though I propose in my original paper that $Q(h)$ ought to be interpreted as what the posterior of h would be if all of e's impact on h were due to its impact on ha as a corporate body.

Proof. Suppose that

$$\frac{P(e|h\neg a)}{P(e|\neg(ha))} = 1 + \epsilon \tag{1}$$

for some ϵ. Then by Bayes' theorem followed by (1),

$$\frac{P^+(h\neg a)}{P^+(\neg(ha))} = \frac{P(e|h\neg a)}{P(e|\neg(ha))} \cdot \frac{P(h\neg a)}{P(\neg(ha))} = (1+\epsilon)\frac{P(h\neg a)}{P(\neg(ha))}.$$

Thus (since $h\neg a \equiv h \wedge \neg(ha)$)

$$P^+(h|\neg(ha)) = (1+\epsilon)P(h|\neg(ha)),$$

or in other words,

$$P^+(h|\neg(ha)) - P(h|\neg(ha)) = \epsilon P(h|\neg(ha)). \tag{2}$$

The following is a theorem of the probability calculus (again because $h\neg a \equiv h \wedge \neg(ha)$):

$$P^+(h) = P^+(ha) + P^+(h|\neg(ha))P^+(\neg(ha))$$

From this theorem and the definition of $Q(h)$,

$$\begin{aligned} P^+(h) - Q(h) &= \left[P^+(h|\neg(ha)) - P(h|\neg(ha))\right] P^+(\neg(ha)) \\ &= \epsilon P(h|\neg(ha))P^+(\neg(ha)) \qquad \text{by (2)} \end{aligned}$$

Since $P(h|\neg(ha))P^+(\neg(ha))$, being the product of two probabilities, is less than or equal to one,

$$P^+(h) - Q(h) \leq \epsilon,$$

as desired. □

References

Fitelson, B. and A. Waterman. (2004). Bayesian confirmation and auxiliary hypotheses revisited: A reply to Strevens. *British Journal for the Philosophy of Science* 56:293–302.

Strevens, M. (2001). The Bayesian treatment of auxiliary hypotheses. *British Journal for the Philosophy of Science* 52:515–538.