Accuracy-first epistemology and scientific progress

Peter Lewis (Dartmouth College)
Don Fallis (Northeastern University)
Branden Fitelson (Northeastern University)

Abstract:

The accuracy-first program attempts to ground epistemology in the norm that one's beliefs should be as accurate as possible, where accuracy is measured using a scoring rule. We argue that considerations of scientific progress suggest that such a monism about epistemic value is untenable. In particular, we argue that counterexamples to the standard scoring rules are ubiquitous in the history of science, and hence that these scoring rules cannot be regarded as a precisification of our intuitive concept of epistemic value.

1. Epistemic value

The accuracy-first program in epistemology is a *monism* about epistemic value: "accuracy … is the sole fundamental source of epistemic value" (Pettigrew 2016, 7). Grounding all epistemic value in accuracy allows the epistemic value of a belief state to be *measured*: the epistemic value of a set of credences is a function of the overall proximity of those credences to the truth. This feature allows some significant results to be proven, including dominance arguments for probabilism (Joyce 1998) and conditionalization (Briggs and Pettigrew 2020), and expected utility maximization arguments for conditionalization (Greaves and Wallace 2006).

However, the *utility* of monism about epistemic value isn't a *justification*. Our goal in this paper is to cast doubt on accuracy monism. Our primary argument is that it cannot give an adequate account of historical cases of *scientific progress*. In making this argument, we assume that science, when conducted properly, is the closest we get to a paradigm of epistemic good practice.

The measure of accuracy we will presuppose in this paper is the *Brier score:* the *total inaccuracy* of credences $\mathbf{c} = (c_1, c_2, \dots c_n)$ in propositions $X = (X_1, X_2, \dots X_n)$ with truth values $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots \omega_n)$ is given by $B(\mathbf{c}, \boldsymbol{\omega}) = \Sigma_i(c_i - \omega_i)^2$. The Brier score is the most popular measure in the literature (Joyce 2009, 275; Pettigrew 2016, 8), but we will mention other measures in passing along the way.

Suppose you have probabilistic credences (1/7, 3/7, 3,7) in a partition of three propositions (A, B, C), where, unknown to you, A is true. Suppose you learn something that conclusively rules out C but is uninformative between A and B, and you conditionalize on this evidence. Intuitively, you are epistemically better off: you have ruled out a false proposition, which decreases the inaccuracy of your beliefs, and learned nothing to distinguish the remaining propositions, which is epistemically neutral. But your initial Brier score is 1.102, and your final Brier score is 1.125. The Brier score says that your epistemic situation has become *worse*, but this seems wrong. The other standard scoring rules—the log rule and the spherical rule—suffer from analogous counterexamples (Lewis and Fallis 2021, 4025). Let us call cases like this *elimination counterexamples*.

Our goal in this paper is to argue that elimination counterexamples are not simply minor clashes with intuition. To make good on this claim, we connect the credence assignments that constitute counterexamples to the Brier rule with actual episodes in the history of science. The elimination counterexamples in the literature are given either in terms of toy examples (Fallis and Lewis 2016, 582) or uninterpreted credences (Dunn 2019, 155). We propose instead to show that elimination counterexamples are ubiquitous in the history of science. That is, it is common to encounter cases of clear epistemic progress in science that the Brier rule counts as the *opposite*. These counterexamples to the Brier rule cannot simply be ignored: to insist on the Brier rule in such cases would do serious harm to our understanding of epistemic progress. Hence they generate a dilemma for the accuracy-first program: either the Brier rule doesn't measure accuracy, or accuracy is not all there is to epistemic value.

2. The Semmelweis case

While working in a maternity ward, Ignaz Semmelweis discovered that hand washing by doctors dramatically reduced the incidence of childbed fever, thereby confirming the hypothesis that disease can be transmitted via dirty hands (Hempel 1966, 3). Semmelweis's discovery was ignored and discounted by the medical establishment, he lost his job, became depressed, and died after a beating at a mental institution at the age of 42 (O'Connor & Weatherall 2019, 77). Only much later did doctors realize the importance of his work, and in the meantime, many women died. It seems clear that behind this human tragedy lies an epistemic one: European doctors' beliefs were *worse* than they should have been.

How can we substantiate the epistemic harm in this case? Initially, it looks easy enough to justify it using a two-hypothesis model. The dominant theory of disease transmission at the time was the *miasma* hypothesis: diseases were atmospheric phenomena caused by "bad air". Semmelweis proposed instead that childbed fever was caused by "cadaveric particles"—particles from corpses carried on the hands of medical students and doctors from the dissection room to the ward. These hypotheses are mutually exclusive, so an increase in credence in the cadaveric particle hypothesis and a corresponding decrease in credence in the miasma hypothesis constitutes epistemic progress.

However, Semmelweis's hypothesis was not, in fact, *true*. As Hempel (1966, 6) points out, it is not particles from corpses *per se* that cause disease; indeed, Semmelweis later obtained evidence that matter from living bodies could sometimes have the same effect. Rather, we now know that it is *microbes* that carry disease, whether from corpses or elsewhere. So let H stand for the proposition that childbed fever is transmitted on hands (rather than via some other medium, such as through the air), and let M stand for the proposition that the disease is carried by microscopic organisms (rather than by cadaveric particles or some other mechanism). Then we can form a partition of *four* (exhaustive, mutually exclusive) hypotheses: H&M, H&~M, ~H&M, ~H&~M. Of these, ~H&~M includes the received view in Semmelweis's time: the hypothesis that the disease is transmitted neither on hands nor via microbes includes the miasma hypothesis. H&~M is the hypothesis that the disease is transmitted on hands, but not via microbes; it includes Semmelweis's cadaveric particle hypothesis. The remaining two possibilities are microbe hypotheses: H&M is the hypothesis that the disease is transmitted on hands by microscopic organisms, and ~H&M is the hypothesis that the disease is transmitted in some other way by microscopic organisms. In the case of childbed fever, H&M is the true hypothesis.

Now consider how Semmelweis's evidence should have affected doctors' credences in these four hypotheses. The evidence that hand washing reduces the incidence of childbed fever is incompatible with hypotheses ~H&M and ~H&~M, so credence in these (false) hypotheses goes to zero. It is consistent with the “hand-borne” hypotheses H&M and H&~M, so credence in these hypotheses goes up, and they remain in the same relative proportion. Of these, H&M is true, and H&~M is false. So credence in a false hypothesis (Semmelweis’s hypothesis) goes up, but so does credence in the true hypothesis.

Does such a credence shift constitute epistemic progress? It is certainly intuitive that it *does*: that is why it was a tragedy that Semmelweis’s colleagues *ignored* his evidence. But what does the Brier score say? Consider a typical doctor in Semmelweis’s time, who is quite confident in the miasma hypothesis compared to the cadaveric particle hypothesis, and for whom the microbe hypothesis barely registers. That is, they assign a fairly low credence to H, the hypothesis that childbed fever is transmitted on hands, and a very low credence to M, the hypothesis that it is carried by microbes. Suppose, then, that their credence in H is 0.2, their credence in M is 0.01, and suppose also that they believe H and M to be independent.[1] These assumptions fix the initial credences of the four hypotheses in the partition, as well as arbitrary disjunctions and negations of these propositions, as shown in Table 1. The total Brier score for the set of propositions before conditionalization is 6.648, and after conditionalization it is 7.840. That is, the Brier score indicates that conditionalizing on Semmelweis’s evidence makes things *worse* from an epistemic perspective. But we know that doctors’ credences would have been *better*, not worse, if they had paid attention to Semmelweis’s evidence. That is, under this assignment of credences, the Semmelweis case is a real-life elimination counterexample to the Brier score.

[1] Nothing hangs on this assumption: it is just a convenient way to fix the initial credences.

Proposition	**Truth value**	**Initial credence**	**Initial Brier score**	**Final credence**	**Final Brier score**
Contradiction	0	0	0.000	0	0.000
H&M	1	0.002	0.996	0.01	0.980
H&~M	0	0.198	0.039	0.99	0.980
~H&M	0	0.008	0.000	0	0.000
~H&~M	0	0.792	0.627	0	0.000
H	1	0.200	0.640	1	0.000
M	1	0.010	0.980	0.01	0.980
H↔M	1	0.794	0.042	0.01	0.980
~(H↔M)	0	0.206	0.042	0.99	0.980
~H	0	0.800	0.640	0	0.000
~M	0	0.990	0.980	0.99	0.980
~(H&M)	0	0.998	0.996	0.99	0.980
~(H&~M)	1	0.802	0.039	0.01	0.980
~(~H&M)	1	0.992	0.000	1	0.000
~(~H&~M)	1	0.208	0.627	1	0.000
Tautology	1	1	0.000	1	0.000
TOTAL			**6.648**		**7.840**

Table 1: Brier scores for a Boolean algebra in the Semmelweis case

3. Generality

We have argued so far that there are plausible credence assignments in the Semmelweis case such that it functions as a real-life counterexample to the Brier score, taken as a measure of epistemic value. But one example is not sufficient to undermine the Brier score as a reasonable precisification of our intuitive concept of epistemic value. We need to show that such counterexamples are *robust*—that there is nothing special about this precise credence assignment. And we need to show that they are *general*—that there is nothing special about the Semmelweis case in particular.

Let us start with generality. In the Semmelweis case, the dominant hypothesis is falsified by some new evidence, and of the remaining hypotheses, a false hypothesis starts with higher credence than the true hypothesis. This looks like quite a commonplace situation. Consider, for example, Foucault's experiment of 1850 to measure the relative speed of light in air and in water. The result was taken to falsify Newton's particle theory of light and confirm Huygens' wave theory. But Huygens' wave theory is not *true*, since light is made up of photons. Hence we have a situation in which a relatively high-credence hypothesis—Newton's particle theory—is ruled out, and another relatively high-credence hypothesis—Huygens' wave theory—is confirmed, but a third, low-credence hypothesis—the photon theory—is true. This is just the sort of situation in which elimination counterexamples to the Brier score arise.

Or consider the Michelson-Morley experiment. The experiment was designed to distinguish between two accounts of the electromagnetic ether: the stationary ether hypothesis, according to which the Earth moves relative to the local ether, and the ether drag hypothesis, according to which the Earth is always stationary relative to the local ether. The result of the experiment was taken to falsify the former. But the ether drag hypothesis is not *true*: the truth is that there is no electromagnetic ether. The true hypothesis had very low credence at the time. Hence this, too, is exactly the kind of situation in which elimination counterexamples to the Brier score arise.

To these examples, it might be objected that physicists in the nineteenth century had *no credence at all* in photon theories, or in theories of light without an ether, since these theories were not formulated until after 1900. But we can formulate a generic version of the photon hypothesis—i.e. the hypothesis that light consists of discrete entities that nevertheless exhibit wave-like properties. Similarly, we can formulate a generic version of the no-ether hypothesis—i.e. the hypothesis that there is no ether and yet light manages to travel through a vacuum.

Nineteenth-century scientists could certainly *consider* such generic hypotheses, but would not have taken them seriously: hence the low initial credence.

Indeed, one might take this kind of example as ubiquitous in science. That is, when an experiment confirms a hypothesis, that hypothesis rarely, if ever, turns out to be simply and absolutely *true*. Rather, it eventually turns out to be *false*, and is replaced by some hypothesis that had very low credence at the time of the original experiment. Indeed, it is a reasonable methodological principle to direct one's experimental efforts toward distinguishing between relatively *plausible* hypotheses. As long as the truth is not one of these initially plausible hypotheses—which is typically the case—then a counterexample to the Brier score will arise whenever one of the plausible hypotheses is eliminated.[2] In this sense, it looks like the Brier score entails that epistemic progress is the *exception* rather than the rule when a false hypothesis is eliminated.

Hence the counterexamples to the Brier rule are quite general. The credences we ascribe in the Semmelweis case are typical of a common situation in the history of science. So counterexamples to the Brier score as a measure of epistemic value are not just a theoretical possibility that can be easily brushed aside.

4. Robustness

Let us turn to the *robustness* of the counterexamples to the Brier score. How sensitive are they to the particular assignment of credences? Fig. 1 shows the extent of counterexample-

[2] Someone who is impressed by the pessimistic meta-induction might hold that we have no reason to think that the true hypothesis ever rises to the level of plausibility. We make no such commitment here: we remain neutral on the force of the pessimistic meta-induction.

producing initial credences for three exhaustive, mutually exclusive hypotheses X_1, X_2, X_3. Here X_1 is the true hypothesis, X_2 and X_3 are false hypotheses, and X_3 is eliminated by new evidence. (In the Semmelweis case, we can identify H&M with X_1, H&~M with X_2, and ~H with X_3.) The horizontal axis indicates initial credence c_1 in hypothesis X_1 and the vertical axis indicates initial credence c_2 in X_2; since the agent's credences are assumed to be probabilistic, credence c_3 is not an independent parameter, but is given by $1 – (c_1 + c_2)$. Since $(c_1 + c_2)$ cannot be greater than 1, the bottom-left triangle (bounded by the two axes and the diagonal) contains all possible probabilistic credences. The shaded area represents initial credences such that elimination of false hypothesis X_3 results in increased inaccuracy according to the Brier score; this is the counterexample region.

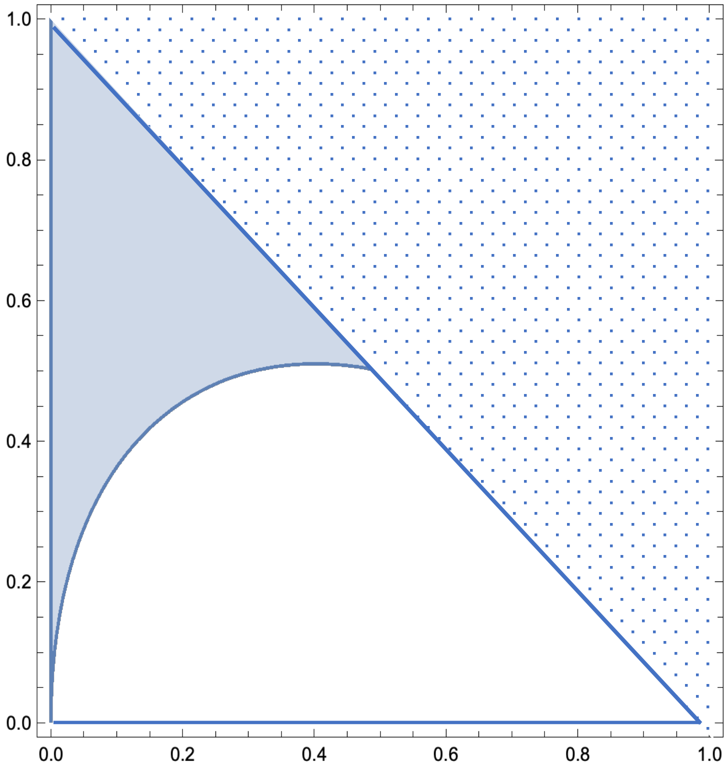

Figure 1: Initial credences producing Brier counterexamples

It is striking that the counterexamples occur over a large, contiguous region of credence-space: it is not just isolated or extremal sets of credences that produce counterexamples to the Brier score. Hence counterexamples are generally robust against small changes in initial credences, and may be robust against quite large changes in initial credences, depending on where in the diagram the counterexample under consideration is located. Indeed, although we don't have the space to do so here, altering the initial credences in the Semmelweis, Foucault, and Michelson-Morley cases shows that reasonable changes in initial credences leave them within the counterexample region.

5. Objections

We have argued that counterexamples to the accuracy-first project are both ubiquitous in the history of science, and robust regarding the exact assignment of priors. However, there are various ways one might object to the claim that the historical episodes we consider constitute counterexamples to the accuracy-first project. First, one might bite the bullet and insist that the historical episodes are *not* instances of epistemic progress—that Semmelweis, and Foucault, and Michelson and Morley, were in fact epistemically *worse off* after their respective experiments. After all, one might argue, conditionalizing on their evidence was the rational policy, insofar as conditionalization maximizes *expected* epistemic value. It just so happens that in each case they got unlucky, so that their *actual* epistemic value decreased.

Our response is that the bullet-biting defense does too much damage to our concept of epistemic value. Even though a precisification of the concept of epistemic value might depart from some of our intuitive judgments, such a wholesale departure from standard usage suggests that we are not really precisifying *epistemic value* at all. The bullet-biting defense claims that epistemic progress in science is often *deferred*—that Semmelweis was in fact initially worse off epistemically, and it was only after further testing that epistemic progress was restored. But this is highly counter-intuitive: surely it would have been better—*epistemically* better—had Semmelweis's contemporaries believed him.

A second line of objection might target our identification of scientific progress with increase in epistemic value. Even if we take good science to be a model of epistemic virtue, it doesn't follow that scientific progress and epistemic progress *coincide*: there might be more to scientific progress than epistemic improvement. Hence an accuracy-first epistemologist might argue that our examples really *are* examples of scientific progress, even though epistemic value decreases.

Admittedly, there *can be* more to scientific progress than increase in epistemic value: for example, science often produces practical benefits as well as knowledge. And this might provide a plausible diagnosis of the Semmelweis case: even though Semmelweis's credences became less accurate, he made scientific progress insofar as he was able to save lives. However, the other cases we have considered are more puzzling. In the Foucault and Michelson-Morley cases there were no immediate practical consequences of the relevant experiments, only epistemic ones, and yet these still seem like clear cases of scientific progress. In fact, the generic nature of elimination counterexamples in science seems to preclude a response in terms of non-epistemic benefit: while *some* such cases might be accompanied by a non-epistemic benefit, there is no reason to think that they *all* will.

A third line of objection might target our identification of epistemic value with *accuracy*. As pluralists about epistemic value, this response seems exactly right to us: in our examples, while *accuracy*, as measured by the Brier score, decreases, *epistemic value*, in the sense relevant to scientific progress, goes up. But this line of objection cannot be used in defense of the accuracy-first program, precisely because this program is committed to monism about epistemic value—to the thesis that accuracy is the *only* fundamental epistemic value.

Nevertheless, there is an approach to epistemic value along these lines that, while not respecting the *letter* of the accuracy-first approach, might give accuracy-firsters most of what they want. That is, even if accuracy is just one epistemic value among several, perhaps an *overall* measure of epistemic value is available, and perhaps this measure can be used to vindicate probabilism and conditionalization without giving the wrong verdict in elimination experiments. We take it that this is the most promising response to the challenge posed by elimination experiments.

6. Scientific progress

To illustrate how this might go, consider the Semmelweis case. Conditionalizing on Semmelweis's evidence leads to a decrease in accuracy according to the Brier score. But one might argue that even though Semmelweis's beliefs get less accurate, nevertheless he made scientific/epistemic progress, and that it is this progress that drives our intuition about the case. After all, Semmelweis was *right* that childbed fever was carried on hands, and it was this insight that allowed him to reduce the incidence of the disease by mandating hand washing.

There are a number of ways one might cash this out, depending on one's preferred account of scientific progress. Let us consider four leading accounts: problem-solving, knowledge, understanding, and verisimilitude (Dellsén 2018). According to the problem-solving account of scientific progress, learning H (i.e. increased credence that childbed fever is carried on hands) allows Semmelweis to solve the problem of the high death rate in the maternity ward. According to the knowledge account, Semmelweis makes progress in that his increase in credence in H means that he now *knows* H: he has acquired a justified true belief in H. According to the understanding account, Semmelweis makes progress in that his increase in credence in H allows him to correctly *explain* the high death rate in the maternity ward. According to the verisimilitude account, Semmelweis's cadaveric particle hypothesis (an instance of H&~M) is closer to the truth (H&M) than the miasma hypothesis (~H&~M), in that it at least gets right that childbed fever is carried on hands. So an increased credence in H amounts to a shift in credence from a hypothesis far from the truth (~H&~M) to a hypothesis closer to the truth (H&~M), thereby increasing verisimilitude.

In all these accounts, scientific progress is associated with an increase in credence in H. That is, one proposition is picked out as particularly relevant to progress. Perhaps, then, we can construct a measure of epistemic value that incorporates a measure of accuracy, but also gives special status to some elements in the algebra of propositions. In what follows we consider combined measures of accuracy and *verisimilitude*, since these have been explored in the literature. We return briefly to other accounts of scientific progress at the end of this section.

Dunn (2019) proposes just such a combined measure of accuracy and verisimilitude. Note that, in the four-hypothesis partition, H&~M and ~H&M are closer to the truth (H&M)

than ~H&~M, since they each get one thing correct. Since we are understanding verisimilitude in terms of the atomic hypotheses H and M and their negations, Dunn argues that these propositions should be accorded special weight in calculating a combined accuracy-verisimilitude score. That is, instead of a straight Brier score over the Boolean algebra of propositions, Dunn proposes a *weighted* Brier score, in which the score for the atomic propositions H and M and their negations is multiplied by a large weight, and the score for all the other propositions is multiplied by a small weight; as before, the epistemic goal is to minimize this score.

We can read the results of this weighted score off Table 1. H and ~H have initial credences of 0.2 and 0.8, respectively, and M and ~M have initial credences of 0.01 and 0.99. If the scores for these propositions get a weight of 1 and the scores for all other propositions get a weight of 0, the weighted Brier score is 3.266. Semmelweis's evidence drives the credences in H and ~H to 1 and 0, respectively, and leaves the credences in M and ~M unchanged at 0.01 and 0.99, for a weighted Brier score of 1.960. According to the weighted Brier score, which takes verisimilitude into account, an agent's beliefs get *better* after incorporating Semmelweis's evidence, as they should. Even if the weighting is not so extreme, the same hopeful result follows. Dunn (2019, 165) recommends non-zero weights for all propositions, so that the resulting weighted Brier score is *proper*: this is important since propriety is a crucial premise in the proofs of probabilism and conditionalization.

This is a promising direction for a defense of something akin to the accuracy-first program: it is not an accuracy monism, but nevertheless provides a unified measure of epistemic value that can ground probabilism and conditionalization. However, obstacles remain. Oddie (2019) argues that any acceptable combined measure of accuracy and verisimilitude must satisfy *proximity*: it must be such that if you redistribute your credences in some false hypotheses so that it is all concentrated on the false hypothesis that is closest to the truth, you do not make things worse according to the measure. Oddie (2019, 576) proves that no measure can satisfy both proximity and propriety, and hence if proximity is accepted as a reasonable constraint, then no combined measure of accuracy and verisimilitude can be used to prove probabilism and conditionalization.

Furthermore, even though Dunn's combined accuracy-verisimilitude measure can defuse some apparent counterexamples from the history of science, others remain. In particular, there are historical cases in which verisimilitude is beside the point. Consider again the example of Foucault's experiment from section 3. If the photon hypothesis is true, neither Newton's particle hypothesis nor Huygens's wave hypothesis is obviously closer to the truth than the other: the photon hypothesis incorporates aspects of both a particulate and a wave theory. More concretely, let P be the hypothesis that light comes in discrete units, and let W be the hypothesis that light obeys a wave equation. Then we can identify Newton's hypothesis with P&~W, Huygens' hypothesis with ~P&W, and the photon hypothesis with P&W. Hence Newton's hypothesis and Huygens' hypothesis are equally verisimilar on a Dunn-style analysis. Consider an agent who initially has equal credence in Newton's hypothesis and Huygens' hypothesis: let us set them both to 0.4.[3] Since, prior to quantum theory, P and W were thought to be incompatible, let us set the typical credence in P conditional on W to 0.01, yielding an

[3] Since Newton's hypothesis was already under threat from diffraction and interference phenomena, a typical scientist might have a lower credence in P&~W than in ~P&W. Nevertheless, neutrality was still presumably rationally permissible.

initial credence in P&W of 0.004. The residual "catch-all" credence in ~P&~W is then 0.196, producing initial credences over the Boolean algebra as shown in column 3 of Table 2. Foucault's evidence eliminates P&~W, yielding the final credences in column 5.

Proposition	**Truth value**	**Initial credence**	**Initial Brier score**	**Final credence**	**Final Brier score**
Contradiction	0	0.000	0.000	0	0.000
P&W	1	0.004	0.992	0.007	0.987
P&~W	0	0.400	0.160	0	0.000
~P&W	0	0.400	0.160	0.667	0.444
~P&~W	0	0.196	0.038	0.327	0.107
P	1	0.404	0.355	0.007	0.987
W	1	0.404	0.355	0.673	0.107
P↔W	1	0.200	0.640	0.333	0.444
~(P↔W)	0	0.800	0.640	0.667	0.444
~P	0	0.596	0.355	0.993	0.987
~W	0	0.596	0.355	0.327	0.107
~(P&W)	0	0.996	0.992	0.993	0.987
~(P&~W)	1	0.600	0.160	1	0.000
~(~P&W)	1	0.600	0.160	0.333	0.444
~(~P&~W)	1	0.804	0.038	0.673	0.107
Tautology	1	1	0.000	1	0.000
TOTAL			**5.400**		**6.152**

Table 2: Brier scores for a Boolean algebra in the Foucault case

We can see from Table 2 that the Brier score over the Boolean algebra increases. This is not surprising: it is an elimination counterexample, as noted in section 3. Furthermore, if we concentrate on P, W, and their negations, the initial Brier score over these propositions is 1.420, and the final Brier score is 2.188. That is, Dunn's measure also indicates that things have gotten worse. Again, this is not surprising: Foucault's evidence shifts credence from P&~W such that most of it goes to a proposition that is equally far from the truth (~P&W), a little of it goes to a proposition that is even further from the truth (~P&~W), and a negligible amount goes to the truth (P&W). Considerations of verisimilitude are of no help here, as the structure of the case initially suggests.[4]

Finally, note that cases like this also challenge other accounts of scientific progress. In the Semmelweis case, H becomes *known*, and this known fact allows Semmelweis to correctly *explain* childbed fever and to correctly solve the *problem* of the high death rate. But in the Foucault case, although credence in W increases somewhat, it can hardly be said to become known, and hence form the basis of correct explanation or problem-solving.[5] Since there are historical cases of this form, an appeal to knowledge, explanation, or problem-solving cannot address *all* elimination counterexamples.

In sum, then, combined verisimilitude-inaccuracy measures conflict with a plausible principle—Oddie's proximity principle—and must also contend with historical counterexamples. These counterexamples also challenge other accounts of scientific progress. Although this is a reasonable direction to look for a defense of something that might do the work of the accuracy first approach, it remains problematic.

7. Explicating epistemic value

The choice of a measure of epistemic value might be posed as a Carnapian explication project: the goal is to construct a measure that is both fruitful and sufficiently similar to our intuitive

[4] Dunn (2019, 162) argues that in cases like this, epistemic value really *does* decrease, because credence becomes concentrated on a single false hypothesis—in this case, on ~P&W. Lewis and Fallis (2021, 4030) reply that, while concentrating credence on a false hypothesis might have *some* relevance to epistemic value, it is implausible that it should carry so much epistemic weight. In any event, Dunn's response here is an instance of the "bullet-biting" defense addressed in section 5.

[5] A Kuhnian approach to scientific progress might divorce problem-solving from truth, but this approach is controversial (Dellsén 2018, 5).

notion.[6] The Brier score is undoubtedly fruitful, but we have argued that it fails on the similarity criterion. In particular, endorsing the Brier score requires us to judge that conditionalizing on Semmelweis's evidence was epistemically negative for the average scientist at the time. What's more, this kind of example is quite generic: endorsing the Brier score requires us to judge that ruling out a false hypothesis is *generally* epistemically negative when the true hypothesis has relatively low credence, and this kind of situation occurs in science quite regularly. This, we maintain, does far too much damage to our intuitive conceptions of epistemic value and scientific progress. The Brier score is so dissimilar from our intuitive judgments that it should not count as an explication of *epistemic value* at all. Either the Brier score fails to measure accuracy, or accuracy is not all there is to epistemic value.

Where does that leave the accuracy-first program? A defender might try to grasp the first horn of the dilemma by devising an alternative measure of accuracy that does not suffer from the problems facing the Brier score. We are skeptical of this approach. Lewis and Fallis (2021, 4031) argue that no measure that obeys reasonable conditions can escape elimination counterexamples altogether, and we suspect that still stronger results are available. That is, we suspect that there is no measure of accuracy that avoids elimination counterexamples and can ground proofs of probabilism and conditionalization.

The second horn of this dilemma, while departing from the monism of the accuracy-first program, offers more hope of defending the general approach. We saw that Dunn's combined measure of accuracy and verisimilitude based on the Brier score can defuse some of the historical counterexamples: even though straight accuracy decreases, this is offset by an

[6] Carnap (1950, 5) also includes exactness and simplicity as desiderata.

increase in verisimilitude, such that the overall epistemic situation in e.g. the Semmelweis case improves. But we have argued that significant historical counterexamples remain, and it is unclear whether a combined measure of accuracy and verisimilitude satisfying the requirements of the accuracy-first program is feasible at all.

The remaining possibility is that the dilemma is inescapable—that there is *no* single measure of epistemic value that both satisfies the foundational assumptions of the accuracy-first program and is sufficiently similar to our intuitive notion. If you have pluralist leanings concerning epistemic value, as we do, this might seem like the natural conclusion, but it would be a serious blow to the accuracy-first program.

References

Briggs, R. A., and Richard Pettigrew (2020), “An accuracy-dominance argument for conditionalization,” *Noûs* 54: 162–181.

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

Dellsén, Finnur (2018), “Scientific progress: Four accounts,” *Philosophy Compass* 13: e12525.

Dunn, Jeffrey (2019), “Accuracy, verisimilitude, and scoring rules,” *Australasian Journal of Philosophy* 97: 151–166.

Fallis, Don, and Peter J. Lewis (2016), “The Brier rule is not a good measure of epistemic utility (and other useful facts about epistemic betterness),” *Australasian Journal of Philosophy* 94: 576–590.

Greaves, Hilary, and David Wallace (2006), “Justifying conditionalization: conditionalization maximizes expected epistemic utility,” *Mind* 115: 607–632.

Hempel, Carl G. (1966), *Philosophy of Natural Science*. Prentice Hall.

Joyce, James M. (1998), “A nonpragmatic vindication of probabilism,” *Philosophy of science* 65: 575-603.

Joyce, James M. (2009), “Accuracy and coherence: prospects for an alethic epistemology of partial belief,” in F. Huber and C. Schmidt-Petri (eds.), *Degrees of Belief*. Springer: 263–297.

Lewis, Peter J., and Don Fallis (2021), “Accuracy, conditionalization, and probabilism,” *Synthese* 198: 4017-4033.

O'Connor, Cailin, and James Owen Weatherall (2019), *The Misinformation Age: How False Beliefs Spread*. Yale University Press.

Oddie, Graham (2019), “What accuracy could not be,” *The British Journal for the Philosophy of Science* 70: 551–580.

Pettigrew, Richard (2016), *Accuracy and the Laws of Credence*. Oxford University Press.