ACCURACY-FIRST EPISTEMOLOGY AND SCIENTIFIC PROGRESS

PETER J. LEWIS
Dartmouth College

DON FALLIS
Northeastern University

BRANDEN FITELSON
Northeastern University

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. The Accuracy Account of Scientific Progress

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. The measure of accuracy we will presuppose in this paper is the *Brier score:* the *total inaccuracy* of credences $\mathbf{c} = (c_1, c_2, \ldots c_n)$ in propositions $\mathbf{X} = (X_1, X_2, \ldots X_n)$ with truth values $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots \omega_n)$ is given by $B(\mathbf{c}, \boldsymbol{\omega}) = \Sigma_i(c_i - \omega_i)^2$, where the credences are real numbers,[1] and the truth values take either the value 0 (false) or 1 (true). Since the Brier score measures inaccuracy, the fundamental epistemic norm according to the accuracy-first program is to minimize the value of the Brier score. This norm allows some significant results to be proven, including dominance arguments for probabilism (Joyce 1998) and conditionalization (Briggs and Pettigrew

[1] Since the Brier score can be used in a proof of the probability axioms, it is not a stipulation of the Brier score that the credences must lie in the range zero through one. Nevertheless, since all the examples we use in this paper involve agents with probabilistic credences, it is safe to assume that all credences lie in this range.

2020), and expected utility maximization arguments for conditionalization (Greaves and Wallace 2006).[2]

However, the *fruitfulness* of accuracy 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. Scientific progress can be characterized in various ways, but these characterizations typically give a central role to an epistemic element: increased *knowledge*, increased *understanding*, or increased *verisimilitude* (Dellsén 2018).[3] Note, though, that we do not assume that scientific progress is *exhausted* by epistemic progress: scientific progress can clearly involve non-epistemic elements, such as facilitating technology and improving human lives. Rather, we assume that scientific progress typically *involves* epistemic progress, and that the epistemic and non-epistemic components of progress in any particular case can usually be distinguished. That is, our interest in cases of scientific progress is not because we *identify* scientific progress with epistemic progress, but because we regard the history of science as a *rich source* of clear cases of epistemic progress.

Let us start by describing a case of scientific progress that does *not* conflict with accuracy monism. We do this in order to illustrate the kind of account of scientific progress we take the accuracy monist to be making. Consider, then, the impact of Harvey's circulatory experiments in the early 1600s on the understanding of the function of the liver (Bolli 2019). According to Galen, the primary function of the liver was to continuously produce blood, which travelled outward through the veins and did not return. By estimating the volume and rate of pumping of the heart, Harvey showed that the weight of blood pumped by a human heart in an hour was four times the average human weight, ruling out the hypothesis that blood is continuously produced by the liver. His experiment did nothing to distinguish between the two other major hypotheses concerning the primary function of the liver, namely that it is the seat of the emotions, and that it is involved in digestion. Nevertheless, even though his experiment didn't reveal the true function of the liver (which is digestive, broadly speaking), it did constitute *progress*: Harvey made scientific progress, and that progress was primarily epistemic.

We can model this experiment as involving belief in three hypotheses—X_1, X_2, X_3—where X_1 is the hypothesis that the primary function of the liver is digestive, X_2 is

[2] The Brier score is the most popular measure of inaccuracy in the literature (Joyce 2009: 275; Pettigrew 2016: 8), but other measures also support these results. We consider the possibility of using a different measure in section 5.

[3] An exception might be Kuhnian problem-solving accounts of scientific progress, which tend to eschew any global sense of epistemic progress. But such accounts remain controversial for that very reason (Dellsén 2018: 5).

the hypothesis that the primary function of the liver is emotional, and X_3 is the hypothesis that the primary function of the liver is blood production. We can take X_1 to be true, and X_2 and X_3 to be false. Consider a scientist in the early 1600s who (quite reasonably) regarded these three hypotheses as equiprobable, since there was no real evidence at the time concerning the function of the liver.[4] That is, their initial credences (c_1, c_2, c_3) are (1/3, 1/3, 1/3). Harvey's evidence eliminates X_3, and is neutral between X_1 and X_2. We assume that the scientist conditionalizes on this evidence, so that their credence in X_3 goes to zero, and their credences in X_1 and X_2 stay in the same ratio, resulting in final credences (1/2, 1/2, 0).[5] We can then compare the scientist's inaccuracy before and after conditionalizing.[6] Their initial Brier score is $(2/3)^2 + (1/3)^2 + (1/3)^2 = 2/3$, and their final Brier score is $(1/2)^2 + (1/2)^2 + 0 = 1/2$. Hence the scientist's inaccuracy has decreased, and accuracy-first epistemology can give an accuracy-based account of their epistemic progress.[7]

Note the structure of this example: we start with a partition (a set of mutually exclusive and exhaustive hypotheses), and the evidence eliminates one of those hypotheses, but is uninformative regarding the others. This elimination constitutes scientific progress. Clearly not all scientific experiments have this structure. Nevertheless, the elimination of a false hypothesis from a partition is a common kind of scientific progress, and for present purposes we consider only examples of this kind.

2. A Counterexample

Let us consider a different example, namely the Michelson-Morley experiment (Michelson and Morley 1887). The experiment was designed to distinguish between two accounts of the electromagnetic ether: the stationary ether hypothesis, according to which the Earth moves through the ether, and the ether drag hypothesis, according to which the Earth drags the ether with it, so that it is always stationary relative to the local ether. The experiment involved measuring the speed of light along two perpendicular directions; unless the Earth is always stationary relative to the local ether, one expects to

[4] We remain neutral in this paper concerning whether the relevant prior credences should be understood purely subjectively, or more objectively as the *rational* ones to hold.

[5] This is an idealization: no hypothesis is ever *absolutely* falsified, for familiar reasons. But it is a harmless idealization for present purposes: the increase in Brier score is not sensitive to whether the credence is reduced *precisely* to zero. Similar comments apply to all our other examples.

[6] See Greaves and Wallace (2006: 615): they describe the comparison in terms of the inaccuracy resulting from the *credal act* of conditionalizing on the evidence, given the true state of the world.

[7] It is common in the accuracy-first literature to consider credences over a Boolean algebra rather than a partition—that is, to include credences in arbitrary disjunctions and negations of elements of the partition. But since we are assuming probabilistic credences, this just has the effect of doubling each Brier score, and hence makes no difference to the comparison of Brier scores.

detect a difference between the two measurements. Since no difference was detected, the result of the experiment was taken to falsify the stationary ether hypothesis. But the ether drag hypothesis is not *true*: the truth is that there is no electromagnetic ether. The Michelson-Morley experiment does not distinguish between the ether drag hypothesis and the no-ether hypothesis. Nevertheless, even though the experiment didn't directly reveal the truth, the elimination of the stationary ether hypothesis constituted scientific progress, where that progress was primarily epistemic.

As before, we can model this experiment as involving belief in three hypotheses—X_1, X_2, X_3—where X_1 is the (true) no-ether hypothesis, X_2 is the (false) ether drag hypothesis, and X_3 is the (false) stationary ether hypothesis. But in this case, a well-informed scientist would not be indifferent between the three hypotheses: since there was no developed theory at the time that dispensed with the electromagnetic ether, the no-ether hypothesis X_1 would be considered a long-shot, and hence would have lower credence than the two ether hypotheses X_2 and X_3.[8] So suppose that our scientist's credences (c_1, c_2, c_3) are initially (0.04, 0.48, 0.48). The null result of the Michelson-Morley experiment eliminates X_3, and credence in X_1, X_2 remain in the same proportions, resulting in final credences of (0.08, 0.92, 0). The initial Brier score is $(0.96)^2 + (0.48)^2 + (0.48)^2 = 1.38$. The final Brier score is $(0.92)^2 + (0.92)^2 + 0 = 1.69$. Note that the Brier score goes *up*: according to the Brier score, the scientist's inaccuracy has increased, and hence, according to accuracy monism, this example constitutes the *opposite* of epistemic progress. But intuitively, the example constitutes clear epistemic progress. Hence we are faced with a prima facie counterexample to the accuracy-first approach.

Our goal in this paper is to argue that elimination counterexamples of this kind are not simply minor clashes with intuition. Elimination counterexamples to the Brier score have been presented in the past, but only 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 score counts as the *opposite*. These counterexamples to the Brier score cannot simply be ignored: to insist on the Brier score 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 score doesn't measure accuracy, or accuracy is not all there is to epistemic value.

A generic response to elimination counterexamples is that even though the scientist's inaccuracy increases after conditionalizing, conditionalization nevertheless maximizes their *expected* accuracy, and hence is the rational thing to do. We do not dispute the provable result that conditionalization maximizes expected accuracy when

[8] One might object that in 1887 there was no no-ether hypothesis that a scientist could attach a credence to. But a scientist in 1887 could certainly *comprehend* the hypothesis that light manages somehow to travel through a vacuum—they would just assign it a low credence.

accuracy is measured by the Brier score (Greaves and Wallace 2006). But note that, according to this response, the scientist suffered an epistemic *mishap*: contrary to expectation, their actual accuracy went down. In other words, according to this response, although they could expect to make epistemic progress on average, in the Michelson and Morley case they unfortunately made the opposite of epistemic progress. This strikes us as a highly counter-intuitive description of the epistemic situation: conditionalizing on the results of the Michelson-Morley experiment did not constitute an epistemic mishap.

3. The Extent of the Problem

We have argued so far that under plausible credence assignments, the Michelson-Morley case 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. So let us consider how *widespread* such counterexamples are.

In Appendix 1, we show how to characterize the extent of counterexample-producing initial credences for three exhaustive, mutually exclusive hypotheses X_1, X_2, X_3 where, as in the examples considered so far, X_1 is the true hypothesis, X_2 and X_3 are false hypotheses, and X_3 is eliminated by new evidence. The result of the calculation is shown in Figure 1. 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. 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.

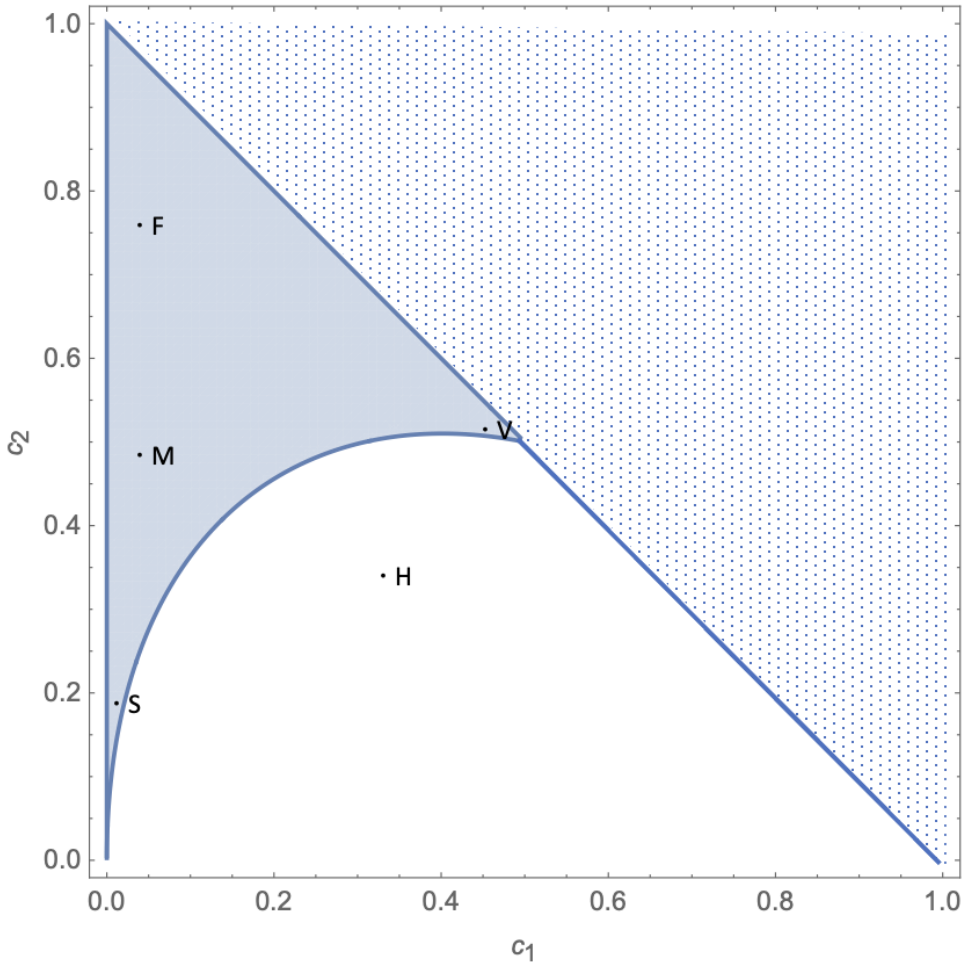

Figure 1: The counterexample region for a three-element partition. (H: Harvey case; M: Michelson-Morley case; F: Foucault case; S: Semmelweis case; V: vaccine-autism case.)

Note also that many different kinds of case fall into the counterexample region. Our initial example of Harvey's experiments, point H on the diagram, falls outside the counterexample region. The Michelson-Morley experiment, point M, falls inside the counterexample region. And it is easy to think of further examples that fall inside the counterexample region. Consider, for example, Foucault's experiment of 1850 to measure the relative speed of light in air and in water (Duhem 1954: 189). 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 actually made up of photons, which are neither Newtonian particles nor Huygens' waves. In this case, since Newton's particle theory of light was already under threat from diffraction and interference phenomena, the credence in Newton's particle theory for a well-informed scientist at the time was already quite low. But their credence in a generic photon hypothesis—a hypothesis according to which light comes in discrete units, and yet exhibits wave-like properties—

would be still lower, since no theory describing such entities had yet been developed. So if X_1 is the (true) photon hypothesis, X_2 is Huygens' (false) wave hypothesis, and X_3 is Newton's (false) particle hypothesis, we might take their initial credences (c_1, c_2, c_3) to be (0.04, 0.76, 0.2), and their final credences to be (0.05, 0.95, 0). The initial credences, point F in fig. 1, fall inside the counterexample region, and indeed the initial Brier score is $(0.96)^2 + (0.76)^2 + (0.2)^2 = 1.54$, and the final Brier score is $(0.95)^2 + (0.95)^2 + 0 = 1.81$. Again, the accuracy-first approach suggests that their epistemic state has become worse, even though this looks like a clear case of epistemic progress.

Or consider Semmelweis's hand-washing experiments in the 1840s. Semmelweis showed that instituting a hand-washing regime for doctors and medical students between the dissection room and the maternity ward led to a dramatic reduction in the incidence of childbed fever (Hempel 1966: 5). Semmelweis took this evidence to rule out the currently dominant hypothesis that childbed fever was caused by a *miasma*, or "bad air", and to confirm Semmelweis's own hypothesis that childbed fever was caused by "cadaveric particles". But Semmelweis's hypothesis was not true: there is nothing about cadavers *per se* that causes childbed fever, and we now know that the cause is microbial. The hypothesis that childbed fever is caused by microscopic life was certainly available in the 1840s, but would have had a very low credence, lower even than Semmelweis's unpopular cadaveric matter hypothesis. So if X_1 is the microbe hypothesis, X_2 is the cadaveric matter hypothesis, and X_3 is the miasma hypothesis, we can take typical initial credences (c_1, c_2, c_3) to be (0.01, 0.19, 0.8), and final credences to be (0.05, 0.95, 0). The initial credences, point S in fig. 1, fall inside the counterexample region: the initial Brier score is $(0.99)^2 + (0.19)^2 + (0.8)^2 = 1.66$, and the final Brier score is $(0.95)^2 + (0.95)^2 + 0 = 1.81$. Again, the accuracy-first approach suggests that conditionalizing on Semmelweis's evidence makes a typical scientist's epistemic state worse, even though it looks like a clear case of epistemic progress.

So far, all the counterexamples have in common that initial credence in the true hypothesis is low. And indeed this seems to be a commonplace situation in the history of science: an experiment distinguishes between the two currently dominant hypotheses, where the truth lies elsewhere. But a glance at the counterexample region shows that low initial credence in the true hypothesis is not necessary for a counterexample. Consider, for example, the epistemic situation concerning the causes of autism in 1999. At that time, it was unknown whether autism is entirely genetic in origin, or whether it has a genetic and an environmental component; more recent twin studies strongly suggest the latter (e.g. Hallmayer et al. 2011). So consider a scientist who assigns a credence of 0.5 to a purely genetic origin, and 0.5 to a combined genetic and environmental origin. Among the possible environmental factors is the MMR vaccine: Wakefield's 1998 study suggested a causal link between MMR vaccination and autism, but the study was small and obviously flawed, so a well-informed scientist would only have a small credence in this particular factor (Godlee, Smith and Marcovitch 2011).

Nevertheless, Wakefield's study had an outsized effect on public opinion, and consequently, studies were rapidly carried out to test Wakefield's hypothesis more rigorously (e.g. Taylor et al. 1999). Such studies showed no link between MMR vaccine and autism.

Consider, then, the following three hypotheses: X_1, that there are environmental causal factors for autism other than MMR vaccination (true); X_2, that there are no environmental causal factors for autism (false); and X_3, that MMR vaccination is a causal factor for autism (false). Suppose a typical well-informed scientist has initial credences (c_1, c_2, c_3) of (0.45, 0.5, 0.05). The studies that conclusively rule out MMR vaccination as a relevant environmental factor produce final credences (0.47, 0.53, 0). The initial credences, point V in fig. 1, fall in the counterexample region: the initial Brier score is $(0.55)^2 + (0.5)^2 + (0.05)^2 = 0.555$, and the final Brier score is $(0.53)^2 + (0.53)^2 + 0 = 0.562$. Ruling out a low-credence hypothesis like this is minor epistemic progress, but it is surely epistemic progress nevertheless; yet the accuracy-first approach entails the opposite.

This last example also addresses an objection one might have to the other counterexamples we have proposed. In the Michelson-Morley, Foucault, and Semmelweis cases, our notional scientist ends up with a credence above 0.9 in the false hypothesis X_2. One might point to this as the *reason* that these cases are not genuine cases of epistemic progress: although the scientist has reduced their credence in one false hypothesis to 0, they have also increased their credence in another false hypothesis to close to 1 (Dunn 2019: 162). We see no reason to conclude that the net effect here is epistemically negative other than a prior commitment to accuracy as the sole measure of epistemic value. But in any event, the vaccine-autism case shows that there are counterexamples in which credence in X_2 barely rises above 0.5. In cases with more than three hypotheses, the final credence in the false hypotheses could be even lower.[9]

Admittedly, though, in *all* our counterexamples, the initial credence distribution over the *false* hypotheses is more evenly spread than the final credence distribution: this is inevitable, given that they are all elimination cases. A case can be made that, keeping credence in the true hypothesis fixed, it is epistemically better to have your credence in the false hypotheses spread evenly rather than unevenly (Dunn 2019: 162; Schoenfield 2022: 393). The Brier score incorporates this feature. Hence, one might argue, none of our counterexamples are cases of genuine epistemic progress, because the increased credence in the true hypothesis is more than counterbalanced by a more uneven distribution of credence in the false hypotheses.

[9] Consider, for example, an eleven-hypothesis case in which the true hypothesis has an initial credence of 0.095, and of the ten false hypotheses, nine have initial credence 0.1, and the tenth has initial credence 0.005. If evidence rules out this last hypothesis, the true hypothesis has final credence 0.0955 and the nine remaining false hypotheses each have credence 0.1005. Then the initial Brier score is 0.90905 and the final Brier score is 0.90907. In this counterexample, no false hypothesis has a final credence much above 0.1.

We do not object to the idea that an even falsity distribution might be an epistemic good, although neither are we committed to it. Our point is just that the accuracy-firster, insofar as they are committed to a proper scoring rule such as the Brier score, is committed to a *particular way* in which the epistemic value of even falsity distribution is to be weighed against the epistemic value of high credence in the truth. That is, one can read our counterexamples as showing that putting *so much* epistemic weight on even falsity distribution flies in the face of our judgments of epistemic progress in science.

Finally, one might object that we are conflating *epistemic* progress with *historical* progress along a path that eventually leads to the truth: science sometimes takes false steps, and the roundabout route to the truth can sometimes include episodes when scientists make the *opposite* of epistemic progress. We do not dispute that this happens. For example, consider Needham and Buffon's 1748 experiment on spontaneous generation (Frost-Arnold 2019: 911): Needham and Buffon boiled gravy in sealed containers to kill all life, and later observed moving microorganisms in the gravy using a microscope. Scientists now think that they probably observed Brownian motion of dead bacteria. This experiment gave support to the false hypothesis that microscopic life-forms spontaneously generate, but also spurred the research that eventually ruled out this hypothesis. In this case, scientists were temporarily *misled* along their historical path to the truth. But note how different this case is from our examples: Needham and Buffon's evidence did not eliminate a false hypothesis, and it did not increase credence in the true hypothesis. That is, although there are undoubtedly historical cases in which scientists were misled by the evidence, our examples are of a very different character. In particular, in our examples, there is clear epistemic progress *in this very episode*.

In sum, there is a wide range of cases of prima facie epistemic progress in the history of science that the accuracy-first approach counts as the opposite. One might quibble with our particular choice of initial credences, but note that the counterexamples are not particularly sensitive to the precise values of the credences—especially in cases like Foucault and Michelson-Morley that lie away from the edges of the counterexample region. One might also worry that in some cases there are other relevant propositions not included in our analysis that might change the accuracy calculation. For example, it may be that the result of Foucault's experiment decreases credence in other false hypotheses concerning light that were part of Newton's particle theory, such as the hypothesis that refraction is caused by attraction between light and the refracting surface. But on the other hand, the result might also *increase* credence in false hypotheses that were part of Huygens' wave theory, such as the hypothesis that space is filled with a medium (ether). Overall, we see no reason to think that such propositions will always exist, or that they will systematically undermine the counterexamples. A further worry might be that in some of our examples (e.g. Semmelweis) it is not clear whether the scientific progress is primarily *epistemic*, since

there also seems to be a good deal of practical progress involved (i.e. saving lives). We feel that there is a clearly separable component of epistemic progress even in the Semmelweis case—Semmelweis was able to save lives *because* he made epistemic progress—but in any event, there are other cases (e.g. Michelson-Morley) in which there is no immediate practical component.

Our general response to all these worries is that elimination counterexamples are *ubiquitous* in the history of science: to paraphrase Laudan (1981: 33), they can be generated *ad nauseam*. So even if some can be ruled out, many others remain. Given the ubiquity of counterexamples to the accuracy-first approach in the history of science, such counterexamples cannot simply be ignored. Consequently, we conclude that the accuracy-first approach to epistemic value is untenable: either the Brier score does not measure accuracy, or accuracy is not all there is to epistemic value. Nevertheless, we think there is an approach to epistemic value that might give accuracy-firsters *almost* everything they want that is worth exploring.

4. Verisimilitude

Consider again the Semmelweis case. One might suspect that the reason Semmelweis made epistemic progress was not because his credences in the various hypotheses he considered became more accurate, but because he shifted his credence from a false hypothesis to another hypothesis that, while still false, was much *closer* to the truth. In particular, his hypothesis that childbed fever is caused by cadaveric particles is much closer to the truth than the hypothesis that it is caused by a miasma. That is, to understand epistemic progress, we need to consider *verisimilitude* as well as accuracy.

Dunn (2019) and Schoenfield (2022) propose just such a combined measure of verisimilitude and accuracy. Note that this approach is *not* an accuracy-first approach: something *other* than accuracy plays a role in epistemic progress, namely verisimilitude. But it might give accuracy-firsters most of what they want, insofar as the combined measure can support arguments for probabilism and conditionalization analogous to those for the Brier score.

To understand the approach, we first need to introduce a more fine-grained set of hypotheses to describe the Semmelweis case. 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.

We can reconstruct the counterexample to the Brier score in terms of this four-element partition. As before, we have a credence of 0.8 in ~H&~M, a credence of 0.19 in H&~M, and a credence of 0.01 in M. Assuming that H and M are (believed to be) independent, we obtain a credence of 0.008 in ~H&M and a credence of 0.002 in H&M. Semmelweis's evidence rules out ~H&M and ~H&~M. The initial and final credences, and their associated Brier scores are shown in Table 1; in fact, the table includes credences and Brier scores for the full Boolean algebra of arbitrary disjunctions and negations of the elements of this partition, since these will be useful in a moment. From Table 1, we see that the Brier score over the four element partition increases from 1.672 to 1.960, and over the whole Boolean algebra it increases from 6.688 to 7.840.

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	0.010	0.980
H&~M	0	0.190	0.036	0.990	0.980
~H&M	0	0.008	0.000	0	0.000
~H&~M	0	0.800	0.640	0	0.000
H	1	0.192	0.652	1	0.000
M	1	0.010	0.980	0.010	0.980
H↔M	1	0.802	0.039	0.010	0.980
~(H↔M)	0	0.198	0.039	0.990	0.980
~H	0	0.808	0.652	0	0.000
~M	0	0.990	0.980	0.990	0.980
~(H&M)	0	0.998	0.996	0.990	0.980
~(H&~M)	1	0.810	0.036	0.010	0.980
~(~H&M)	1	0.992	0.000	1	0.000
~(~H&~M)	1	0.200	0.640	1	0.000
Tautology	1	1	0.000	1	0.000
TOTAL			**6.688**		**7.840**

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

Now note that, in the four-hypothesis partition, H&~M and ~H&M are closer to the truth (H&M) than is ~H&~M, since they each get one thing correct. In particular, Semmelweis's hypothesis (H&~M) is closer to the truth than the miasma hypothesis

(~H&~M), since it at least gets right that childbed fever is transmitted on *hands*. Hence the credence shift we see in this case, from ~H&~M to H&~M, is from a hypothesis that is further from the truth to a hypothesis that is closer to the truth, even though both are false. Since we are understanding verisimilitude in terms of the atomic hypotheses H and M and their negations, Dunn and Schoenfield (following Greaves and Wallace 2006: 628) argue 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, they propose 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 0.192 and 0.808, 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 3.264. 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 *close* 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.[10] However, even though a combined accuracy-verisimilitude measure can defuse the Semmelweis counterexample, its applicability to the other counterexamples is problematic. In some cases, this is because the logical structure that Dunn and Schoenfield rely on is absent. Consider the Michelson-Morley experiment, for example. We might try to reconstruct the partition in this case in terms of two propositions: that the ether *exists* (rather than not), and that the ether is *dragged* (rather than not). But in this case, if the ether does not exist, then the question of whether the ether is dragged or not doesn't arise. Hence a Dunn-Schoenfield analysis of verisimilitude is unavailable in this case. Similar comments apply to the vaccine-autism case: if there are no environmental causes of

[10] Schoenfield (2022: 375) argues that the incorporation of verisimilitude is not a *departure* from accuracy-first epistemology, since the weights just determine how much one cares about the accuracy of a given proposition. We note that the weights themselves are not given by accuracy considerations. But we do not need to take a side on whether the approach is a departure from accuracy-first epistemology or a variant of it.

autism, then the question of whether there are *vaccine-linked* causes in particular doesn't arise.

But even when the appropriate logical structure is present, applying the Dunn-Schoenfield technique doesn't always resolve the counterexample. Consider again the example of Foucault's experiment. 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 particle and a wave theory. More concretely, let D 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 D&~W, Huygens' hypothesis with ~D&W, and the photon hypothesis with D&W; additionally, we have a "neither wave nor particle" hypothesis ~D&~W. On this analysis, Newton's hypothesis and Huygens' hypothesis are equally verisimilar.

We can reconstruct the counterexample to the Brier score in terms of this four-element hypothesis as shown in Table 2. As before, we set credence in Newton's hypothesis D&~W to 0.2, and credence in Huygens' hypothesis ~D&W to 0.76, and we divide the remainder of the credence equally between the true hypothesis D&W and the false hypothesis ~D&~W (since neither was part of a developed theory at the time). The result is an initial Brier score (over the entire Boolean algebra) of 6.312, and a final Brier score of 7.420: as expected, the accuracy-first approach implies that the epistemic situation has become worse. But now suppose we incorporate verisimilitude by weighting the propositions D, W, and their negations. If these propositions have *all* the weight, then the initial Brier score is 1.312 and the final Brier score is 1.904: the weighted measure *still* says that the epistemic situation has become worse. A less extreme weighting does not change this qualitative result. So the verisimilitude strategy does not resolve this case satisfactorily. And this is not surprising: a glance at Table 2 shows that the main effect of conditionalizing on Foucault's evidence is to shift credence from D&~W to ~D&W, which are equally far from the truth (D&W).

Proposition	Truth value	Initial credence	Initial Brier score	Final credence	Final Brier score
Contradiction	0	0.000	0.000	0	0.000
D&W	1	0.020	0.960	0.025	0.951
D&~W	0	0.200	0.040	0	0.000
~D&W	0	0.760	0.578	0.950	0.903
~D&~W	0	0.020	0.000	0.025	0.001
D	1	0.220	0.608	0.025	0.951
W	1	0.780	0.048	0.975	0.001
D↔W	1	0.040	0.922	0.050	0.903
~(D↔W)	0	0.960	0.922	0.950	0.903
~D	0	0.780	0.608	0.975	0.951

~W	0	0.220	0.048	0.025	0.001
~(D&W)	0	0.980	0.960	0.975	0.951
~(D&~W)	1	0.800	0.040	1	0.000
~(~D&W)	1	0.240	0.578	0.050	0.903
~(~D&~W)	1	0.980	0.000	0.975	0.001
Tautology	1	1	0.000	1	0.000
TOTAL			**6.312**		**7.420**

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

5. 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 notion.[11] 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 count many clear-cut cases of epistemic progress in the history of science as the opposite. This, we maintain, does far too much damage to our intuitive conception of epistemic 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 a combined measure of accuracy and verisimilitude based on the Brier score can defuse some historical counterexamples: even though straight accuracy decreases, this is offset by an increase in verisimilitude, such that the overall epistemic situation in e.g. the Semmelweis case improves. But we have argued that many historical

[11] Carnap (1950: 5) also includes exactness and simplicity as desiderata.

counterexamples remain: some resist analysis via this method, and others do not involve an increase in verisimilitude.

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.

Appendix: Calculations Underlying Figure 1

In our examples, we have a partition of three hypotheses: $\{X_1, X_2, X_3\}$, the initial credences of which are given by $\{c_1, c_2, c_3\}$, respectively. Here X_1 is the true hypothesis, X_2 and X_3 are false hypotheses, and X_3 is eliminated by new evidence. Because the three hypotheses form a partition, we have the following two background constraints.

(1) $\{c_1, c_2, c_3\} \in [0,1]$,

(2) $c_1 + c_2 + c_3 = 1$, i.e., $c_3 = 1 - (c_1 + c_2)$

Assuming conditionalization, the posterior probabilities of $\{X_1, X_2, X_3\}$ — upon learning that X_3 is false — are given by $\left\{\frac{c_1}{c_1+c_2}, \frac{c_2}{c_1+c_2}, 0\right\}$, respectively. The Brier Score of the initial credence function is given by:

(3) $(c_1 - 1)^2 + (c_2 - 0)^2 + (c_3 - 0)^2 = (c_1 - 1)^2 + {c_2}^2 + (1 - (c_1 + c_2))^2$

The Brier Score of the posterior credence function is given by:

(4) $\left(\frac{c_1}{c_1+c_2} - 1\right)^2 + \left(\frac{c_2}{c_1+c_2}\right)^2 + (0 - 0)^2 = \left(\frac{c_1}{c_1+c_2} - 1\right)^2 + \left(\frac{c_2}{c_1+c_2}\right)^2$

Therefore, the posterior credence function is more inaccurate than the initial credence function iff (4) > (3), *i.e.*, iff

(5) $\left(\frac{c_1}{c_1+c_2} - 1\right)^2 + \left(\frac{c_2}{c_1+c_2}\right)^2 > (c_1 - 1)^2 + {c_2}^2 + (1 - (c_1 + c_2))^2$

We can use *Mathematica*'s **RegionPlot** function to plot the region of $\{c_1, c_2\}$-space in which constraints (1), (2), and (5) obtain (i.e., the region of initial credal space containing the Brier counterexamples): the result is Figure 1. We can also use *Mathematica* to give a closed-form expression for this region (albeit one containing a **Root** object).[12]

Acknowledgments

The authors would like to thank Richard Pettigrew, Ben Levinstein, and audiences at the Northeastern Epistemology Workshop (Boston, November 2021) and the Philosophy of

[12] A *Mathematica* notebook containing all the relevant calculations (and generated figures) can be downloaded from the following URL: http://fitelson.org/sp_appendix.nb.

Science Association Biennial Meeting (Pittsburgh, November 2022), for helpful discussions of earlier versions of this paper.

References

Bolli, Roberto (2019). William Harvey and the Discovery of the Circulation of the Blood, Part II. *Circulation Research, 124*(9), 1300–1302.

Briggs, R. A., and Richard Pettigrew (2020). An Accuracy-Dominance Argument for Conditionalization. *Noûs, 54*(1), 162–181.

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

Dellsén, Finnur (2018). Scientific Progress: Four Accounts. *Philosophy Compass, 13*(11), e12525.

Duhem, Pierre (1954). *The Aim and Structure of Physical Theory* (Philip P. Wiener, Trans.). Princeton University Press.

Dunn, Jeffrey (2019). Accuracy, Verisimilitude, and Scoring Rules. *Australasian Journal of Philosophy, 97*(1), 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*(3), 576–590.

Frost-Arnold, Greg (2019). How to Be a Historically Motivated Antirealist: The Problem of Misleading Evidence. *Philosophy of Science, 86*(5), 906–917.

Godlee, Fiona, Jane Smith, and Harvey Marcovitch (2011). Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent. *British Medical Journal, 342*, c7452.

Greaves, Hilary, and David Wallace (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind, 115*(459), 607–632.

Hallmayer, Joachim, Sue Cleveland, Andrea Torres, et al. (2011). Genetic Heritability and Shared Environmental Factors among Twin Pairs with Autism. *Archives of General Psychiatry, 68*(11), 1095–1102.

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

Joyce, James M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science, 65*(4), 575-603.

Joyce, James M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Franz Huber and Christoph Schmidt-Petri (Eds.), *Degrees of Belief* (263–297). Springer.

Laudan, Larry (1981). A Confutation of Convergent Realism. *Philosophy of Science, 48*(1), 19–49.

Lewis, Peter J., and Don Fallis (2021). Accuracy, Conditionalization, and Probabilism. *Synthese, 198*, 4017–4033.

Michelson, Albert A., and Edward W. Morley (1887). On the Relative Motion of the Earth and the Luminiferous Ether. *American Journal of Science*, *34*(203), 333–345.
Pettigrew, Richard (2016). *Accuracy and the Laws of Credence*. Oxford University Press.
Schoenfield, Miriam (2022). Accuracy and Verisimilitude: The Good, the Bad, and the Ugly. *The British Journal for the Philosophy of Science,* 73(2), 373–406.
Taylor, Brent, Elizabeth Miller, C. Paddy Farrington, Maria-Christina Petropoulos, Isabelle Favot-Mayaud I, Jun Li, and Pauline A. Waight (1999). Autism and Measles, Mumps, and Rubella Vaccine: No Epidemiological Evidence for a Causal Association. *The Lancet 353*(9169), 2026–2029.