Preferring the less reliable method

Wolfgang Schwarz

Compare the following two ways of responding to the weather report's "probability of rain" announcement.

> *Good:* Upon hearing that the probability of rain is x, you come to believe to degree x that it will rain.
>
> *Bad:* Upon hearing that the probability of rain is x, you become certain that it will rain if x > 0.5, otherwise certain that it won't rain.

The Bad process seems bad, not just because it may lead to bad decisions. It seems *epistemically bad* to respond to a "70% probability of rain" announcement by becoming absolutely certain that it will rain. The resulting attitude would be unjustified and irrational.

Can we explain the comparative badness of the Bad process solely by appeal to the overriding epistemic goal of truth? It seems not.

Let's compare the objective reliability of the two methods. Roughly speaking, the reliability of a belief-generating process measures its tendency to generate true beliefs. How do we apply this to the Good process which typically generates only partial beliefs? We don't want to count an 0.3 belief that it will rain as true iff it will rain. A natural move is to measure reliability in terms of the distance between the degree of belief and the truth value of the relevant proposition (1 for true, 0 for false). So a reliable process would tend to generate high degrees of belief in true propositions, and low degrees in false propositions. Call the distance between degree of belief and actual truth value the belief's *inaccuracy*.

Using this measure, it turns out that under normal circumstances, the Bad process is *more reliable* than the Good process. To illustrate, consider 100 occasions on which the weather forecast announces a 70 percent probability of rain. Let's also assume that in 70 of these cases, the announcement is followed by rain, and in 30 it isn't. (It is a pretty good weather forecast.) The Bad method would make you certain that it rains on each occasion, so your distance from the truth is 0 in 70 cases and 1 in 30. Average inaccuracy: 0.3. The Good method would give you a belief of degree 0.7 that it will rain, so your distance from the truth is 0.3 in 70 cases and 0.7 in 30 cases. Average inaccuracy: 0.42. On average, the Bad process brings you closer to the truth.

Note that as a counterexample to reliabilism, this is quite different from thought-experiments involving far-fetched scenarios where the reliability of a process comes apart from (what we take to be) its actual reliability. Even under perfectly normal and common conditions, the Bad process is more reliable than the Good one.

The above reasoning also shows that the Bad process has lower expected inaccuracy that the Good method from the point of view of an agent who uses the *Good* method. If you follow the Good method and hear the "70% chance of rain" announcement, you can calculate that the expected inaccuracy of the Bad method is 0.3, while the expected inaccuracy of your own method is 0.42. (If you follow the Bad method, you'll judge the Bad method to have 0 expected inaccuracy, compared to 0.3 for the Good method.)

This is odd. Here we have two methods; we know that one of them is more reliable, that it is more likely to bring us closer to the truth; but still we think it would be irrational to use it. We judge that, from a purely epistemic perspective, one ought to use the other, less reliable method.

A while ago, I claimed that the appearance of truth as the unifying epistemic virtue might be an illusion based on the fact that any belief-forming method whatsoever will automatically be regarded as truth-conducive by agents who use it. It looks like we have a counterexample to this as well. At least it is not true that any belief-forming method automatically appears more accuracy-conducive than its rivals to agents who use it.

Four quick comments.

First, the present puzzle is closely related to the puzzle discussed in Allan Gibbard's "Rational Credence and the Value of Truth" (2008), who should therefore get all the credit. Gibbard assumes that it is irrational to have degrees of belief which, by their own lights, are less accurate than certain other degrees of belief; and he argues that this requirement of rationality cannot be explained solely on the assumption that rational belief "aims at truth", although it can be explained on the assumption that rational belief aims at successful action.

Second, it is tempting to argue against the Bad method by considering its long-run accuracy. If you follow the Bad method and otherwise update by conditioning, you may easily end up with more inaccurate beliefs than if you follow the Good method. But if truth is your goal, why not combine a deviant response to weather forecasts with a deviant update process? It is easy to show that some such combinations have higher reliability, and higher expected long-run accuracy than the sensible combination of the Good method with conditioning (see Gibbard's follow-up note "Aiming at Truth over Time" (2008), page 6).

Third, another problem with the Bad method is that it doesn't generalise well. Suppose the weather forecast says that there's a 30 percent chance of rain, a 40 percent chance of sunshine, and a 30 percent chance of neither rain nor sunshine. And suppose you apply the Bad method not just to the rain statement, but also to the two others. You'd end up being a) certain that it won't rain, b) certain that the sun won't shine, and c) certain that it will either rain or the sun will shine. But arguably, this is an impossible state of mind; the attitude ascriptions (a)-(c) are inconsistent. So if one could show that the Bad method, even restricted to rain forecasts, may (easily) lead to impossibilities like this, that would solve the puzzle. I don't think this can be shown. But another possibility is that the compensatory methods you have to use in addition to the Bad method in order to restore consistency will have a cost in expected accuracy. And then maybe any consistent package of methods containing the Bad process would end up less accuracy-conducive than the reasonable package containing the Good process. That would be nice.

(A lot of people think that there is nothing inconsistent about the attitude *ascriptions* (a)-(c), even though there is something inconsistent about the ascribed attitudes. We would then have to ask whether the rationality requirement of having consistent attitudes can be explained solely by appeal to the goal of truth or accuracy; see e.g. Jim Joyce's "Accuracy and coherence: Prospects for an alethic epistemology of partial belief" (2009) for the latest moves in this game. The short answer is that it can't be done. But as I said, I don't find this very problematic, because I'm sympathetic to Ramsey's view that the requirements of probabilistic coherence are analytic and therefore not in need of epistemic defense.)

Fourth, I've measured inaccuracy simply as the distance between belief and truth value. But there are other measures. If we use squared distance, the problem disappears. There may even be good reasons for using this measure, apart from solving the present problem (Joyce mentions some at the end of the paper just cited). But none of these reasons seem to be based on truth as the overriding epistemic goal. If *all that matters* for epistemic rationality is closeness to the truth, it is not clear why closeness should be measured by squared distance rather than absolute distance.

On the other hand, it is also not clear why closeness should be measured by absolute distance. So I might have to qualify the claim in the other blog post: on one disambiguation of "accuracy-conduciveness" there are clear counterexamples to the hypothesis that epistemic goodness is a matter of accuracy-conduciveness. On this reading, accuracy (or truth) therefore doesn't appear to

be the overriding epistemic goal. On another disambiguation, there is such an appearance, but it is an illusion based on the fact that any belief-forming method whatsoever automatically appears accuracy-conducive by agents who use it.