THE FALLACY OF CALIBRATIONISM

YOAAV ISAACS

1. Introduction

How should an agent respond to information about the reliability of her judgments? Various philosophers have argued for various versions of calibrationism, a view according to which an agent's credences should correspond to the (suitably defined) expected reliabilities of her judgments. Calibrationism gives intuitively reasonable verdicts, and it applies straightforwardly even when an agent is worried that her judgments may be flawed. Because of these advantages, even philosophers who don't want to endorse calibrationism in full generality are often inclined to endorse its verdicts in a wide array of cases.[1] But calibrationism is misguided. Calibrationism relies on the base-rate fallacy, a classic mistake in probabilistic epistemology. Thus while calibrationism is intuitive, it cannot be correct.

2. A Puzzle About Reliability

Consider the following puzzle:

> **Hung Over**: Moishe is a *shochet*. He is responsible for ritually slaughtering animals in accordance with Jewish dietary laws. The very first thing that Moishe does when someone brings him an animal is determine whether or not it is *kosher*, as only some animals may be consumed by observant Jews. Slaughtering a kosher animal would provide his community with food, while slaughtering a non-kosher animal would render his tools ritually unclean. Ordinarily it is very easy for Moishe to distinguish kosher from non-kosher animals. But on this occasion Moishe is hung over. The previous night he was celebrating *Purim*, and in accordance with custom got very, very drunk. Moishe knows that when he is hung over he correctly identifies whether or not an animal is kosher only 50% of the time. When he is hung over he might

[1] This leads to a puzzle about why calibrationism—if false—delivers so many plausible verdicts. See Schoenfield (2015) for more.

> mistake a cow for a pig or a pig for a cow. On this occasion Moishe judges that the animal he is being asked to slaughter is a cow, and thus is kosher. How confident should Moishe be that his judgment is correct?

It's very natural to think that Moishe should be 50% confident that his judgment is correct. After all, Moishe knows that he's hung over, and further knows that when he's hung over his judgments are right only 50% of the time. Sure, sometimes Moishe is lucky when he's hung over, but it's equally true that sometimes Moishe is unlucky when he's hung over. Calibrating his credence to his known reliability seems entirely sensible. This line of thinking can be generalized to contexts in which the reliability of an agent's judgment is uncertain.

> **Calibrationism Schema**: If the expected reliability of an agent's judgment regarding p is r, then if that agent judges that p the agent's credence in p should be r.

Some philosophers have endorsed such calibrationist thinking, and even philosophers who don't want to endorse calibrationism in full generality are often inclined to endorse its verdicts in a wide array of cases.[2] Calibrationism seems like a sensible position, so it's natural that philosophers are attracted to it.

3. Two Precisifications

There is an intuitive connection between expected reliability and rational credence. There are, however, various possible precisifications of expected reliability, and thus various precisifications of calibrationism. Calibrationism faces different difficulties depending on its precisification so it's important to get clear on just what sort of calibrationism one is analyzing. The most natural precisification of calibrationism is fallacious; a somewhat less natural precisification is trivial.

To aid our analysis, let's unpack Moishe's story a little more. Let's say that when Moishe is hung over and is judging whether or not an animal is kosher a tiny *dybbuk*[3] in Moishe's head flips a fair coin one side of which is stamped "kosher" and the other side of which is stamped "non-kosher". The dybbuk makes Moishe judge according to the outcome of the coin flip, so whether or not the animal in front of Moishe is kosher, he will judge its status correctly 50% of the time. We'll assume that Moishe knows the structural features of his reliability.

[2]See White (2009), Sliwa & Horowitz (2015), and Schoenfield (2015). The calibrationist thinking in these papers will be explored soon.

[3]A malicious, possessing spirit from Jewish mythology.

3.1. **A Fallacious Precisification.** Moishe knows that whether or not the animal in front of him is kosher he will judge its status correctly 50% of the time. For now, let's take this fact to entail that Moishe knows that he's 50% reliable. This is a very natural conclusion to draw. (I'll say more about this notion of reliability and about alternative notions of it later.)

Moishe knows that he's 50% reliable at judging whether or not animals are kosher. Moishe knows he judged that the animal that was brought to him is kosher. Should Moishe thereby be 50% confident that animal in front of him is kosher?

No. Moishe's judgment may not give him evidence one way or the other, but it's not the only thing that matters. Quite apart from Moishe's judgment it's substantially more likely than not that the animal brought to him is kosher. Kosher animals are routinely brought to shochets while non-kosher animals are virtually never brought to shochets. What kind of person brings a non-kosher animal to a shochet?

Reasoning that Moishe's known 50% reliability means that he should have 50% credence in his judgment is natural, but it is incorrect. This pattern of reasoning commits a classic epistemological mistake: the base-rate fallacy.[4] Most simply, the base-rate fallacy involves erroneously ignoring the significance of prior probabilities. Before Moishe even saw the animal it was hugely more probable that it would be kosher than that it would be non-kosher. That prior probability matters, and it makes it the case that Moishe should be more confident than not that the animal is kosher. Ignoring the initial probabilities of hypotheses is one of the classic mistakes in probabilistic epistemology. It is known as the base-rate fallacy because it can be thought of as ignoring differences in base-rates—the base-rate of kosher animals brought to shochets being different than the base-rate of non-kosher animals brought to shochets. Reliability matters, but it is not the only thing that matters. Prior probability matters too.

Note that probabilistic reasoning untouched by the base-rate fallacy does not validate a tendentious credulity in one's judgments. Moishe judged that the animal was kosher, and he should be very confident that his judgment was correct. But if Moishe had judged that the animal was non-kosher then he should have been very confident that his judgment was incorrect. Initially, Moishe is 50% confident in the correctness of the judgment he is about to make. But when he makes his judgment and thus learns what it is that should change his credence

[4]For the classic explanation of the erroneous reasoning underlying the base-rate fallacy, see Kahneman & Tversky (1973).

that his judgment is correct. If he learns that he judged 'kosher' that shouldn't make him more confident that the animal is kosher. But he was already very confident that the animal would be kosher, so he should be more confident than not that he that judged correctly. If he learns that he judged 'non-kosher' that shouldn't make him more confident that the animal is non-kosher. But he was already very confident that the animal wouldn't be non-kosher, so he should be more confident than not that he that judged incorrectly.

Imagine an extreme case, one in which the prior probability of an animal being non-kosher is 0. Suppose Moishe is visiting a cow farm, and is certain that there are no pigs for miles. No matter what judgments Moishe makes in his addled state he should still be certain that the animals he sees are cows rather than pigs. It would be entirely unreasonable for Moishe to be guaranteed to adopt 50 / 50 credence about whether each animal he sees is a cow or a pig. Knowing that one is bad at distinguishing cows from pigs when one is hung over should not make one suddenly worried that cow farms are overrun with pigs. Moishe can still know that his judgments will be wrong 50% of the time: He knows that half the time he will judge that some animal on the cow farm is a cow, and that those judgments will all be correct. He knows that half the time he will judge that some animal on the cow farm is a pig, and that those judgments will all be incorrect.

One might worry that Moishe has extra information in these situations that affects his expected reliability, and thus licences asymmetric credences. Maybe Moishe should think his judgment is more than 50% reliable when the animal he is evaluating was antecedently likely to be kosher and he judges that it is kosher. It's a plausible-seeming thought, but it's no help here. We stipulated a notion of expected reliability according to which Moishe's knowledge about the coin-flip structure of his judgments entails that Moishe knows that he's 50% reliable. Whatever additional information Moishe has cannot affect his expected reliability. This precisification of calibrationism is committed to an untenable pattern of reasoning.

There are other notions of expected reliability that calibrationism might employ—indeed, we'll look at an alternative notion shortly. But this notion of expected reliability renders calibrationism fallacious.

3.2. **A Trivial Precisification.** Moishe knows that whether or not the animal in front of him is kosher he will judge its status correctly 50% of the time. That need not mean that the expected reliability of any judgment of Moishe's is 50%. So what else could it mean?

Here's a fairly natural idea: The expected reliability of a judgment is the credence that it's appropriate for an agent to have that the judgment is correct in light of that agent's total evidence. The expected reliability of one of Moishe's upcoming judgments will still be 50%. But (given the high prior probability that the animal is kosher) the expected reliability of Moishe's judgment that the animal is kosher will be greater than 50%. A precisification of calibrationism using this notion of expected reliability will not be fallacious. But all is not well.

The main problem is that this notion of expected reliability makes calibrationism trivial. Regarding this case calibrationism would only mean that however confident Moishe should be that his judgment that the animal is kosher is correct, that's how confident he should be that the animal is kosher. But that's trivial—it's obvious that the judgment that the animal is kosher is correct if and only if the animal is kosher.

As it is a trivial theory, this notion of calibrationism is unsatisfyingly unhelpful. It offers no purchase on what the rational probabilities are. The expected reliabilities of Moishe's judgments can vary wildly; the general structure of his judgments will not pin down the expected reliabilities for particular circumstances. It's even unnatural to talk about "expected" reliabilities in this context; the probability that a judgment is correct seems more like its reliability than like its expected reliability.

There are other notions of expected reliability that calibrationism might employ—indeed, we looked at an alternative notion earlier. But this notion of expected reliability renders calibrationism trivial.[5]

4. Calibrationist Narratives

Now that we have a sense of what a connection between expected reliability and rational credence can amount to, let's look at some of the narratives that have been used to motivate calibrationist thinking.

Roger White first systematized calibrationist thinking, and White (2009) is the seminal text of the calibrationist literature. To his credit, White is explicitly aware that prior probabilities matter, and the bulk of his analysis is formally impeccable. But the base-rate fallacy can sneak up on anyone. White considers the following case:

> [Calibrating] can seem like a bit of common sense. I have a long track record of getting about 90% of my

[5]An even more trivial precisification is also possible. One might think that an agent judges that p whenever the agent has credence greater than .5 in p, and think that her expected reliability is her credence that her judgment is correct. Since it's obvious that a judgment that p is correct if and only if p is true, it is trivial that an agent's credence in p will match her credence that a judgment that p is correct.

> answers right on arithmetic tests. Unless I know I'm enjoying the benefits of a mind-enhancing elixir, surely I shouldn't expect a better success rate on the test I've just taken. You arbitrarily point to Q. 57. How confident am I that this answer is correct? Well, it seems right to me. But I can't very well have more than 90% confidence that it is right without judging likewise for each of the questions, can I? But if I have more than 90% confidence in the truth of each of my answers, then I should expect that I got more than 90% of them right. For no apparent reason I can think of, I would be supposing that I have suddenly started batting above my average. That this strikes us as foolish is a reason to think my confidence that my answer to Q. 57 (or any other question) is right should be constrained to 90%. Since I know that my answer to the question is p, my confidence that my answer is correct must equal my confidence that p. So we reach the conclusion that as the Calibration Rule insists, my confidence in p should equal my expected reliability of 90%.[6]

White's case makes some odd stipulations about how his mathematical reasoning works. Notably, it's quite odd to think that his arithmetical phenomenology is completely uniform—that regarding each question one answer seems right and the rest seem wrong and that there's nothing more to the story. It's very natural to have some questions seem harder than others, to have some answers seem more compelling than others.[7] But even granting White's odd stipulations his analysis is incorrect.

For ease, let's think of the arithmetic problems at stake in this case as multiple choice. (Fill in the blank problems are tantamount to multiple choice, just with more choices.) We can imagine White scrutinizing his available answers to each question, attempting some arithmetic, and then having one of the answers seem to be right. Let's further suppose that White's prior probabilities are such that he'll always be more confident in the answer that seems right than he is in any of the alternatives (this seems to follow the case as White articulates it). Even given all this, the only way for a 90% overall reliability to mandate credence .9 in the answer that seems to be right is for White

[6]White (2009).

[7]For a thorough—and not entirely negative—analysis of such epistemic seemings, see Hawthorne & Lasonen-Aarnio (MS).

to initially think that all the possible answers are equiprobable.[8] If an initially plausible answer seems right, White should have higher credence than .9. If an initially implausible answer seems right, White should have lower credence than .9. The average confidence White has in his answers should work out to around .9, but there is absolutely nothing that mandates that value in any particular case. Any credence between 0 and 1 is rationalizable.[9] White may be right that calibrating seems like common sense, but in that case common sense is mistaken.

Inspired by White (2009), Paulina Sliwa and Sophie Horowitz advocate calibrationism. They consider the following case:

> **Calculation**: Anton is an anesthesiologist, trying to determine which dosage of pain medication is best for his patient: A or B. To figure this out, Anton assesses some fairly complex medical evidence. When evaluated correctly, this kind of evidence determines which dose is right for the patient. After thinking hard about the evidence, Anton becomes highly confident that dose B is right. In fact, Anton has reasoned correctly; his evidence strongly supports that B is the correct dose.
>
> Then Sam, the chef at the hospital's cafeteria, rushes in. "Don't administer that drug just yet," he says guiltily. "You're not in a position to properly assess that medical evidence. I slipped some reason-distorting mushrooms into your frittata earlier as a prank. These mushrooms make you much less reliable at determining which dose the evidence supports: in the circumstances you presently face—evaluating this type of medical evidence, under the influence of my mushrooms—doctors like you only tend to prescribe the right dose 60% of the time!" In fact, Sam is mistaken: the mushrooms he used were just regular dried porcini, and Anton's reasoning is not impaired in the least. But neither he nor Anton knows (nor has reason to suspect) this.[10]

They write,

[8]That is, assuming that White's reliability is uniform. Without an assumption of uniform reliability calibrationism faces even more problems. More on this soon.

[9]White must think that the answer he gives is more plausible than any of the others, so his credence that his answer is right must be at least $\frac{1}{n}$, where n is the number of possible answers. The more possible answers there are, the lower White's credence in the correctness of his answer can be.

[10]Sliwa & Horowitz (2015).

> Consider the following line of thought: upon learning about the mushrooms, Anton should become less confident that B is the right dose. How much should Anton reduce confidence? We said that Sam told Anton that doctors in his position only prescribed the right dose 60% of the time. So a natural view to take is that after hearing Sam's testimony, Anton should only be .6 confident that dose B is right. Moreover, you might think, Anton should be certain that his total evidence supports .6 confidence in B. . . . [A]fter hearing Sam's testimony, Anton should be .6 confident that dose B is right. He should be .6 confident because he expects to be 60% reliable; he should be .6 confident in dose B because that's what his first-order evidence actually supports.[11]

Again, the judgment that Anton should have credence .6 that dose B is correct because he is 60% reliable commits the base-rate fallacy. This judgment erroneously ignores prior probabilities. If it was antecedently more likely that B was the correct dose than that A was, then Anton should have credence greater than .6 that dose B is correct after he updates on his evidence. If it was antecedently more likely that A was the correct dose than that B was, then Anton should have credence less than .6 that dose B is correct after he updates on his evidence. Only in the special case in which dose B and dose A were initially equally likely to be the correct dose should Anton have credence .6 that dose B is correct after he updates on his evidence.[12] Prior probabilities must be reckoned with.

Miriam Schoenfield is perhaps the most prominent critic of calibrationism. Even so, I will argue that she is not quite as critical of the position as one ought to be. While Schoenfield does not endorse calibrationism in full generality, she is explicitly sympathetic to calibrationist reasoning. She writes, "The basic idea is that we should not think of calibrationism as a principle of epistemic rationality. . . Rather, we should think of calibrationism as a principle of reasoning."[13] And as a principle of reasoning, "calibrationism looks like a promising candidate."[14] She considers the following case:

[11]Sliwa & Horowitz (2015).

[12]That is, assuming that Anton's reliability is uniform. Without an assumption of uniform reliability calibrationism faces even more problems. More on this soon.

[13]Schoenfield (2015).

[14]Schoenfield (2015).

> **Hypoxia**: You are a pilot flying to Hawaii. You suddenly realize that you might not have enough fuel to get there. If you don't have enough fuel, you should land right away. Fortunately, you have evidence that entails either that you have enough fuel to get to Hawaii (H) or that you don't (¬H). You do some calculations and breathe a sigh of relief. It looks like you'll make it. You then get a message from ground control: "Dear pilot, you are flying at an altitude that puts you at great risk for hypoxia, a condition that impairs your ability to reason properly. People flying at your altitude, who do the sorts of calculations you just did, only reach the conclusions entailed by their evidence 50% of the time."[15]

Schoenfield writes,

> It seems intuitive to many people that, in **Hypoxia**, you should only have a 0.5 credence in H. But why 0.5? It's very tempting to think that the reason you should have a 0.5 credence in H has something to do with the fact that you learned that you are 50% reliable in these circumstances. Indeed, if you think you should have a 0.5 credence in H in the case above, you probably also think that if, instead, you'd learned that you were 55% reliable, your credence should be 0.55, if you learned that you were 60% reliable, 0.6, and so on. So an attractive way of explaining the judgment in **Hypoxia** appeals to a bridge principle connecting your expected degree of reliability to the credence that it is rational for you to adopt.[16]

Schoenfield means for calibrationism to deliver these verdicts and generalize them.

As we have seen, calibrationism untenably ignores prior probabilities. Suppose that you're initially 99.999% confident that you have enough fuel to make it to Hawaii. (One tends not to attempt such a flight unless one is quite confident that one is not going to run out of fuel mid-ocean.) Suppose that you attempt to calculate whether or not you have enough fuel, and given ambient circumstances you take yourself to be 50% reliable at performing that calculation. For convenience, let us suppose that such 50% reliability is uniform.[17] That is,

[15]Schoenfield (2015).

[16]Schoenfield (2015).

[17]I will discuss non-uniform reliabilities shortly.

if you have enough fuel you expect to calculate that you have enough fuel 50% of the time and to calculate that you don't have enough fuel 50% of the time, and if you don't have enough fuel you expect to calculate that you have enough fuel 50% of the time and to calculate that you don't have enough fuel 50% of the time. Suppose you calculate that you have enough fuel—how should you update your credences? It would be completely unreasonable for you to adopt credence .5 that you have enough fuel and credence .5 that you don't have enough fuel. The calculation cannot inevitably provide such terrible news.[18] Given your 50% reliability your calculation cannot provide any news at all. A 50% reliability is a total lack of correlation. Note that a fair coin is a 50% reliable indicator of whether or not it will rain tomorrow, whether or not God exists, whether or not the moon is made of cheese, and absolutely anything else. Stipulating that the coin landing Heads indicates that the proposition in question is true and that the coin landing Tails indicates that the proposition is false, the coin can get absolutely any proposition right 50% of the time. Given your 50% reliability your calculation does not provide any evidence, and thus you should remain 99.999% confident that you have enough fuel to make it to Hawaii. And a 55% reliability—though not irrelevant—should certainly not guarantee a precipitous drop in credence either.[19] A 55% reliable calculation that you have enough fuel to get to Hawaii should make you slightly more confident that you have enough fuel to get to Hawaii. A 55% reliable calculation that you don't have enough fuel to get to Hawaii should make you slightly less confident that you have enough fuel to get to Hawaii. Either way you should remain well above 99.99% confident that you have enough fuel to get to Hawaii. Calibrationism ignores the significance of your initial 99.999% confidence that you have enough fuel. Thus Calibrationism is false.

The calibrationist narratives presented by White, Sliwa and Horowitz, and Schoenfield all presuppose that the probability that a judgment is correct can be read off of the reliability of the faculty of judgment without regard to the prior probability of what is judged. And as we have seen that is simply incorrect. You cannot automatically have credence .5 that any particular judgment made by a 50% reliable shochet is correct. You cannot automatically have credence .9 that that any particular arithmetic calculation performed by a 90% reliable arithmetic calculator is correct. You cannot automatically have credence .6

[18]Given 50% reliability, a calculation that you don't have enough fuel would have the identical effect.

[19]To .55 or .45, depending on the outcome of the calculation.

that any particular dosage calculation made by a 60% reliable dosage calculator is correct. And you cannot automatically have credence .5 that any particular fuel calculation performed by a 50% reliable fuel calculator is correct.

It is wrong to ignore the prior probabilities of the outcomes in question, and it is wrong to ignore the possibility of non-uniformly reliable processes. The pattern of reasoning that White, Sliwa and Horowitz, and Schoenfield endorse is fallacious.

5. Prospects for Calibrationist Intuitions

We've seen that the most natural precisification of calibrationism is not viable regarding a judgment that p when the prior probability of p does not equal .5.[20] One might therefore think that calibrationism should simply be restricted to judgments about propositions with prior probability .5. But even regarding such propositions, calibrationism routinely goes astray.

So far, we have been assuming that the expected reliabilities of judgments are uniform. Given a uniform reliability, an agent whose judgment regarding p has expected reliability of $\frac{x}{100}$ will judge correctly $x\%$ of the time whether or not p is true. In such a case, if p has prior probability .5 then a judgment that p will licence credence $\frac{x}{100}$ in p.

But what if an agent's expected reliability is non-uniform? What if an agent's judgment regarding p is more likely to be correct if p is true than if it isn't, or more likely to be correct if p isn't true than if it is? (These cases are like the dybbuk / coin toss case discussed earlier, but with a slightly more complex structure.)[21] There are lots of ways to have an expected reliability of $\frac{x}{100}$. The expected reliability of an agent's judgment regarding p is equal to the weighted average of her expected reliabilities given p and given $\neg p$. More formally, it is—

$$\Pr(\text{agent judges that } p|p)\Pr(p) + \Pr(\text{agent judges that } \neg p|\neg p)\Pr(\neg p)$$

Assuming that p has prior probability .5, any intermediate expected reliability can be produced by one uniform expected reliability and by uncountably infinitely many non-uniform expected reliabilities.

[20] Assuming that the expected reliability of the judgment is neither 0 nor 1.

[21] For a toy model of non-uniform reliability, one can imagine that the dybbuk in Moishe's head initiates different chance processes depending on whether or not the animal in front of Moishe is kosher. For example, suppose that if the animal is kosher then the dybbuk rolls a 6-sided die 5 sides of which say "kosher" and 1 side of which says "non-kosher", and if the animal is non-kosher then the dybbuk rolls a 6-sided die 4 sides of which say "kosher" and 2 sides of which say "non-kosher".

Let's look more closely at the sort of difference that a non-uniform expected reliability can make. Let's assume that p has prior probability .5 and that the expected reliability of the agent's judgment regarding p is .55.[22] What credence should an agent have on the basis of her judgment that p? That will depend on the structure of her expected reliability regarding p. At one extreme, the agent is at her most reliable when p is true at at her least reliable when p is false. When p is true she always judges that p, and when p is false she still judges that p 90% of the time. Under these conditions, the agent's judgment that p should lead her to adopt credence $\frac{50}{95}$ in p. At the other extreme, the agent is at her least reliable when p is true at at her most reliable when p is false. When p is true she judges that p 10% of the time, and when p is false she never judges that p. Under these conditions, the agent's judgment that p should lead her to adopt credence 1 in p.[23] Any credence between $\frac{50}{95}$ and 1 can be rationalized by some structure of expected reliability. The 55% rationalized by a uniform expected reliability is only one of uncountably infinitely many possibilities.

The lesson is not that this precisification of calibrationism applies only to judgments regarding propositions with prior probability .5 and uniform expected reliability. The lesson is that this precisification of calibrationism is a fallacious line of reasoning that leads one astray in nearly all cases, albeit not in some extremely rare cases in which it is accidentally correct.

6. An Overview of the Dialectic

It is less than entirely clear what calibrationism is. But the most natural ways to precisify it are beset by severe problems.

If calibrationism deals with the initial probability that an agent will judge correctly[24], if the reliabilities in question are broadly of the form $\Pr(\text{I judge that } p \mid p)$, then calibrationism commits the base-rate fallacy. If calibrationism deals with the probability that some specific judgment is correct, if the reliabilities in question are broadly of the form $\Pr(p \mid \text{I judge that } p)$, then calibrationism is trivial. One can respond to a charge of fallaciousness by articulating a calibrationism that is not fallacious—but that calibrationism will be trivial. One can respond to a charge of triviality by articulating a calibrationism

[22]For these purposes, the case of expected reliability .5 is not propitious.

[23]A helpful heuristic: the former condition makes the agent's judgment that p maximally likely and thus minimally significant, whereas the latter condition makes the agent's judgment that p minimally likely and thus maximally significant.

[24]In White (2009), the objective chance of an agent judging correctly plays this role.

that is not trivial—but that calibrationism will be fallacious. Such responses cannot save calibrationism. It is very hard to extract a principle that is neither fallacious nor trivial from the extant treatments of calibrationism. If the expected reliability of a judgment that p and the expected reliability of a judgment that ¬p are guaranteed to be the same, it's likely because calibrationism doesn't take everything relevant into account and is thereby fallacious. If the expected reliability of a judgment that p and the expected reliability of a judgment that ¬p are not guaranteed to be the same, it's likely because calibrationism takes everything relevant into account and is thereby trivial. Calibrationism can avoid any problem, but it cannot avoid all problems.

Note that other sorts of principles—principles which need neither be fallacious nor trivial—may better express the sensibilities behind calibrationism. For example, Schoenfield articulates a principle, Independence, which requires one's credences to be "independent of the reasoning in question."[25] It is not entirely clear what Independence amounts to.[26] Two interpretations seems reasonably plausible. (1) Independence might mean that agents should look only to general facts about the reliability of their judgments when evaluating the credibility of a particular judgment. In this case, Independence commits the base-rate fallacy. (2) Independence might mean that agents should bracket certain portions of their evidence.[27] Yet whether or not such bracketing is appropriate, if Independence only requires bracketing then there is no important connection between expected reliability and rational credence of the sort that defines calibrationist thinking. It is the connection between expected reliability and rational credence that must either be fallacious or trivial.

7. Life Without Calibrationism

Calibrationism gets some cases wrong.[28] Does it also get some cases right? How many of calibrationism's intuitive verdicts can be justified?

We've seen that calibrationism goes awry when the prior probability of the proposition being judged is not .5. We've seen that even when the prior probability of the proposition being judged is .5, calibrationism goes awry when the expected reliability of a judgment is not uniform. But let us suppose that the prior probability of the proposition being judged is .5 and that the expected reliability of the judgment is uniform.

[25]Schoenfield (2015)

[26]This sort of principle was originally proposed in Christensen (2011).

[27]For more about bracketing see Elga (2007).

[28]Here I consider the most natural notion of calibrationism, which is fallacious.

Does calibrationism deliver sensible verdicts in such circumstances? And if so, do those sensible verdicts require calibrationism?

We can answer these questions. When the prior probability of the proposition being judged is .5 and the expected reliability of a judgment is uniform then calibrationism does deliver sensible verdicts, and those sensible verdicts do not require calibrationism. Under such circumstances calibrationism provably delivers the verdicts of standard probabilistic epistemology.

By Bayes' theorem, the following relationship holds between probabilities concerning a proposition and probabilities concerning a judgment about that proposition:

$$Pr(p \mid \text{judgment that } p) = \frac{Pr(p)\, Pr(\text{judgment that } p \mid p)}{Pr(p)Pr(\text{judgment that } p \mid p) + Pr(\neg p)Pr(\text{judgment that } p \mid \neg p)}$$

We can apply this formula directly. Given that the judgment has uniform expected reliability r, $Pr(\text{judgment that } p \mid p) = r$ and $Pr(\text{judgment that } p \mid \neg p) = (1 - r)$. Given that p has probability .5, the equation thus reduces to:

$$Pr(p \mid \text{judgment that } p) = \frac{.5\, r}{.5r + .5(1 - r)} = \frac{.5\, r}{.5} = r$$

By Bayes' theorem, when the prior probability of a proposition is .5 and the expected reliability of a judgment regarding that proposition is uniform, the probability of the proposition given the judgment is equal to the expected reliability of the judgment.

This result helps explain the intuitive appeal of calibrationism. It makes sense that there should be some context in which the rational credence corresponds to the expected reliability of a judgment. This result shows that not only is there such a context, it is a particularly natural context. There's an intuitive sense in which .5 more natural than any other probability, and there's an intuitive sense in which a uniform expected reliability is more natural than any non-uniform expected reliability. Calibrationism gets the right answer in maximally natural cases. But maximally natural cases are vanishingly rare, and so calibrationism is untenable.

Standard probabilistic epistemology corresponds to calibrationism when calibrationism seems to work well, and it does not correspond to calibrationism when calibrationism goes astray. Extant cases thus do not suggest that standard probabilistic epistemology needs revision. And if standard probabilistic epistemology does not need revision then it is revised at our peril.

8. The Role of Base-Rate Reasoning

I have argued that the most natural precisification of calibrationism commits the base-rate fallacy. Moishe knows that in his addled state his judgments about whether or not an animal is kosher are only 50% reliable. But Moishe also knows that nearly all the animals brought to him (99%, let's say) are kosher. So when he judges that an animal is kosher he should not lower his credence that the animal is kosher to .5. Instead, his credence should remain at .99. Since Moishe's judgment is uncorrelated with whether or not the animal is kosher, Moishe's credence should remain at the expected base rate.

One might worry that this analysis misrepresents what Moishe's judgments are.[29] One might worry that Moishe is only unreliable with respect to judgments based on a subset of his evidence, but that he is not unreliable with respect to judgments based on all his evidence. Perhaps Moishe's judgments based only on his visual evidence are 50% reliable—half the time cows and pigs look the way they should and half the time cows look like pigs and pigs look like cows—but Moishe's judgments based on all his evidence (including his evidence about base-rates) are more reliable. Alternatively, perhaps Moishe's judgments based on all his evidence really are unreliable—he doesn't just struggle to distinguish cows from pigs visually but also struggles to apply base-rate reasoning. In that case, one might feel that Moishe really should only have credence .5 that some animal in front of him is kosher whatever the base rate, on the grounds that it makes sense for Moishe to ignore information that he can't trust himself to think through appropriately.

I note a few things in response. While we did not stipulate what a judgment is, we did stipulate that Moishe knows that his judgments are 50% reliable. Thus the expected reliability of Moishe's judgment is guaranteed to be .5. Such judgments need not be based on Moishe's visual faculties in any special way, and indeed may may be based on the totality of Moishe's evidence. Suppose that when judging a proposition Moishe thinks about everything he can—his experiences, his information about base-rates, and anything else that seems germane—and after this process has gone on for a while it either outputs a judgment that the proposition is true or a judgment that the proposition is false. Even so, Moishe might be certain that he's so inept at this process that the judgments produced by it are uncorrelated with the propositions being judged. The output of a process that is uncorrelated with a proposition is irrelevant to the appropriate credence for that proposition,

[29] I thank an anonymous referee for raising this worry.

so Moishe's credence should not go to .5 as calibrationism says, but should instead remain unchanged. Even if judgments are based on his base-rate information, that does not mean that the expected reliabilities of those judgments are a guide to rational credences. The prior probabilities of the propositions being judged have a specific role in determining the posterior probabilities of those propositions, and it is fallacious to ignore that role even if prior probabilities are given some other role in affecting which propositions get judged to be true.

Here's a way of seeing that prior probabilities have a role which is distinct from expected reliabilities. Calibrationism gives a single number for an agent's expected reliability concerning a proposition—the judgment that p is true and the judgment that $\neg p$ is true each fall under the agent's expected reliability regarding p. But there is no guarantee that the appropriate credence to have in p conditional on the judgment that p is true is the same as the appropriate credence to have in $\neg p$ conditional on the judgment that $\neg p$ is true. Prior probabilities routinely mandate an asymmetry between those two conditional probabilities—the prior probabilities of the two propositions may be different, and the structure of the agent's expected reliability may be different depending on which proposition is true. A single number—the expected reliability of the agent's judgment—just doesn't provide enough information to determine rational credence, and that's why it's fallacious to conform credences to expected reliabilities alone. And it won't do to fine-grain calibrationism and characterize the expected reliability of a judgment that p separately from the expected reliability of a judgment that $\neg p$, as such a fine-grained calibrationism will be trivial. The expected reliability of a judgment that p is the probability that the judgment that p is correct, and is thus straightforwardly also the probability that p is true.[30] The structure that calibrationism needs to avoid being fallacious would make it trivial.[31]

[30]In the context of calibrationism it's assumed that when an agent judges that p the agent is certain that they judge that p. Opacity about judgment-contents—which could drive a wedge between the agent's credence in p and the agent's credence that their judgment that p is correct—is thus ruled out.

[31]An anonymous referee notes a further reason not to match one's credences to the expected reliability of one's judgment: judgments are based on all of one's evidence, yet it might be that the expected reliability of a restricted judgment (based on only some of one's evidence) could have greater expected reliability. Perhaps an agent knows herself to be especially poor at dealing with some particular sort of evidence, such that she's more reliable when she puts that sort of evidence out of her mind. In such cases it is indeed plausible that an agent would be better off calibrating her credences to her restricted judgments than she would be calibrating

Prior probabilities have ineliminable significance in standard probabilistic epistemology. There is thus an important sense in which the bracketing of base-rate information (as is perhaps mandated by Independence principles) has little structural effect. Base-rate information is meant to shape relevant prior probabilities. Moishe knows that the base-rate of kosher animals presented to him is 99%, so absent further information Moishe's prior probability that an animal about to be presented to him will be kosher is 99%. But if Moishe didn't know that base-rate, or were required to ignore his knowledge of that base-rate then he'd need a prior probability nonetheless. And there's absolutely no guarantee that in the absence of information the prior probability of every proposition is .5—that's clearly probabilistically incoherent. Ignoring prior probabilities is committing the base-rate fallacy; it doesn't matter whether those prior probabilities are based on base-rate information or not. Prior probabilities can easily problematize calibrationism even in the absence of base-rate information—it's the particular prior probabilities that pose a problem for calibrationism, not where those prior probabilities come from. And prior probabilities aren't just crucial for standard probabilistic epistemology, they're also crucial for calibrationism: the expected reliabilities in terms of which calibrationism makes its recommendations are determined by prior probabilities.

According to standard probabilistic epistemology an agent is rationally required to update her credences by conditionalization. This rational requirement does not presuppose that agents are perfectly reliable at conditionalization. If an agent has a hard time updating her credences by conditionalization, that just means that she has a hard time updating her credences rationally. Being unreliable at doing what is rationally required doesn't change what's rationally required. It may seem unpalatable that the requirements of rationality should be such that even an agent trying her best to follow them may not be able to do so reliably, but this fact is unavoidable. It's true that agents cannot always be reliable conditionalizers, but it is equally true that agents cannot always be reliable calibraters. Hangovers, fallibility, reason-distorting mushrooms, and hypoxia could just as easily be stipulated

---

her credences to her judgments; the expected accuracy of her credences is straightforwardly superior that way. However, calibrating her credences to her restricted judgments would still be committing the base-rate fallacy. And the expected accuracy of conditionalizing is better than the expected accuracy of calibration to any sort of judgment. For more, see Greaves & Wallace (2005).

to make an agent unreliable at calibrating.[32] It is therefore hard to see sufficient reason to say that agents who are very bad at taking prior probabilities into account are not rationally required to. And it's even harder to see sufficient reason to say that agents who are only slightly imperfect at taking prior probabilities into account are not rationally required to.

Calibrationism is of limited viability even given dubiously favorable stipulations. Consider an agent who cannot reliably conditionalize. If she tries to conditionalize then she knows that 50% of the time she will adopt the credence mandated by conditionalization, c, and 50% of the time she will adopt $(1 - c)$. She can, however, reliably calibrate her credences to her expected reliability at conditionalizing, and adopt credence .5 in a proposition if she so chooses.[33] Suppose that we disallow the view that she ought to conditionalize—the only options we're considering for her are trying to conditionalize (whatever its consequences may be) and calibrating. Given these stipulations, what should she do? Since the orthodoxy in probabilistic epistemology is that she should conditionalize it is far from clear what she should do when that verdict has been disallowed. But there is a very reasonable sense in which it is better for her to calibrate than it is for her try to conditionalize. The upside of the credence she gets if she tries to conditionalize and succeeds is plausibly outweighed by the downside of the credence she gets if she tries to conditionalize and fails.[34] But note that even in this already-strained example calibrationism is viable only accidentally. If the agent were instead 90% reliable at conditionalizing it might make no sense whatsoever for her to adopt credence .9 in a proposition, as .9 might be a worse credence for her to adopt than either c or $(1 - c)$. She might be better off trying to conditionalize than she would be calibrating. And if she were allowed to just adopt some middling credence like .52, that might be better still. Even if conditionalization is disallowed, calibrationism should not take its place.

Calibrationism needs the expected reliability of judgments to be definable in a way that does not straightforwardly reduce to rational credence yet which still plausibly constrains rational credence. But calibrationist narratives do not hint at such a definition of expected

[32]There are strong arguments that ordinary agents are imperfectly reliable at following any epistemic procedure. For more, see Williamson (2008) and Hawthorne & Srinivasan (2013).

[33]Note that this rendering of calibrationism concerns the expected reliability of conditionalizing rather than the expected reliability of judgments.

[34]This follows for any strictly proper scoring rule. For more, see Joyce (1998).

reliability. There is no such definition of expected reliability. The core idea of calibrationism comes from reading reliabilities off of relative frequencies—Moishe gets 50% of the judgments right so he's 50% reliable, Roger White gets 90% of the math problems right so he's 90% reliable, doctors in Anton's position get 60% of diagnoses right so they're 60% reliable, and pilots with hypoxia get 50% of fuel calculations right so they're 50% reliable. But tying credences to such relative frequencies is fallacious. If one conditionalizes then one doesn't commit that fallacy. And even if one can't conditionalize, one still shouldn't commit that fallacy.

9. Conclusion

Calibrationism is an intuitive and appealing position. But it is demonstrably wrong. Calibrationism is intuitive and appealing because it is an instance of the base-rate fallacy, and the base-rate fallacy is intuitive and appealing. Less intuitive and less appealing versions of calibrationism that do not commit the base-rate fallacy are possible, but these fall prey to triviality. It remains to be seen if some theory in the broad spirit of calibrationism is called for. But calibrationism itself is a mistake.[35]

[35]For helpful feedback, I thank John Hawthorne and Alan Hájek. This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

References

Christensen, David (2011). "Disagreement, Question-Begging, and Epistemic Self-Criticism". *Philosophers' Imprint* 11.

Elga, Adam (2007). "Reflection and disagreement". *Noûs* 41 (3):478502.

Greaves, Hilary & Wallace, David (2005). "Justifying conditionalization: Conditionalization maximizes expected epistemic utility". *Mind* 115 (459): 607-632.

Hawthorne, John & Lasonen-Aarnio, Maria. "Not So Phenomenal!". MS.

Hawthorne, John & Srinivasan, Amia. 2013. *Disagreement Without Transparency: Some Bleak Thoughts*. In Christensen, David and Lackey, Jennifer. 2013. *The Epistemology of Disagreement: New Essays*. Oxford University Press.

Joyce, James M. (1998). "A nonpragmatic vindication of probabilism". *Philosophy of Science* 65 (4): 575-603.

Kahneman, Daniel & Tversky, Amos. (1973). "On the psychology of prediction". *Psychological Review* 80 (4):237-251.

Joyce, James M. (1998). "A nonpragmatic vindication of probabilism". *Philosophy of Science* 65 (4): 575-603.

Schoenfield, Miriam (2015). "A Dilemma for Calibrationism." *Philosophy and Phenomenological Research* 91 (2):425-455.

Sliwa, Paulina & Horowitz, Sophie. (2015). "Respecting *all* the evidence". *Philosophical Studies* 172 (11):2835-2858.

White, Roger. (2009). "On Treating Oneself and Others as Thermometers". *Episteme* 6 (3):233-250.

Williamson, Timothy. 2008. Why epistemology cannot be operationalized. In *Epistemology: New Essays*, ed. Quentin Smith, pp. 277-300.