ROGER WHITE

ON TREATING ONESELF AND OTHERS AS THERMOMETERS

ABSTRACT

I treat you as a thermometer when I use your belief states as more or less reliable indicators of the facts. Should I treat *myself* in a parallel way? Should I think of the outputs of my faculties and yours as like the readings of two thermometers the way a third party would? I explore some of the difficulties in answering these questions. If I am to treat myself as well as others as thermometers in this way, it would appear that I cannot reasonably trust my own convictions over yours unless I have antecedent reason to suppose that I am more likely than you to get things right. I appeal to some probabilistic considerations to suggest that our predicament as thermometers might not actually be as bad as it seems.

1. THERMOMETERS AND AGENTS

The thermometer outside tells me it's 60° out. Insofar as I take it to be a reliable indicator of the temperature, I take this as a reason to suppose that it is 60° degrees out. Of course I don't believe just any reading from the thermometer. If it says that it is −10° degrees in the middle of July, I'll conclude that it is malfunctioning. In general, I factor in the thermometer's reading with whatever other evidence I have bearing on the temperature.

I often treat *you* as a thermometer in this way. I take your beliefs on certain matters as indicators of how things are. In a similar way my deference to your opinion depends on my estimate of you as an indicator on the matter and is tempered by whatever other considerations I have bearing on it. The much discussed case of *peer disagreement* is just a special case of taking the opinions of others as evidence. Now if it is appropriate for me to treat you as a thermometer in this way, then it would appear that I deserve the same treatment, not just from you but from myself. I am an agent like you, whose judgments are a more or less reliable indicator of the facts. Surely what information I have about my own reliability as an indicator should be factored into my judgment in a parallel way to your reliability. The natural picture is that each of us carries his own thermometer in his pocket (each of us carries his own brain in his skull). If our thermometers give conflicting readings, it seems that I can't reasonably put more confidence in the reading mine gives unless I *antecedently* had reason to suppose that mine is more likely to give a

DOI: 10.3366/E1742360009000689

correct reading. I can't very well think, "As my thermometer says, it's 60° out. So since yours says it is 65°, it must be wrong." If we shouldn't think this way about indicators we carry in our pockets, why should it be appropriate for the indicators we carry in our heads?

This picture – call it the thermometer model – appears to have radically skeptical consequences. I'm often in no position to judge that you are any less likely than me to arrive at the truth on a matter, and we often disagree. It appears that I ought to be very skeptical of much of what I believe (and the same goes for you, I'm afraid). The question of to what extent the thermometer model is appropriate when thinking about disagreeing agents (especially when one of them is oneself) is perplexing. I have nothing definitive to offer here. I'll begin by muddling my way through the issue, hopefully casting doubt on the thermometer model but also showing how no position on the matter is entirely satisfying. I'll be seeing whether thinking of disagreement in terms of Bayesian updating helps us settle the matter. (Spoiler alert!) It doesn't. It turns out, though, that even accepting the thermometer model doesn't have to lead to such skeptical consequences as we might have thought. I'll explain why.

2. PEERAGE

Epistemic peers are often taken to be agents who share all their evidence and score equally well on some measure of epistemic virtue, which might include factors like intelligence, intellectual honesty, and diligence (see Kelly 2005). It seems appropriate to factor in further variables that affect one's cognitive performance on an occasion: whether an agent is alert, distracted, or stoned. The list could go on, but it is clear that what these factors have in common is that they are relevant to judging in advance how likely it is that an agent will form a correct opinion on some matter. For our purposes we could lump these together under a measure of the agent's degree of *reliability* at answering a question correctly. I suggest we think of this along the lines that we think about objective *chance*. Given my cognitive capacities, what evidence I have to go on, and various environmental factors, I have a certain propensity to form a true belief as to whether *p*. This notion of reliability is a modal one, being logically independent of any actual performance. A reliable thermometer has an objective propensity to give accurate readings, even if by some unfortunate fluke it rarely does (conversely a hopelessly unreliable thermometer might by chance give accurate readings). But there is a strong connection between information of reliability and rational confidence:

> **Reliability-Credence**: Conditional on the reliability of a process producing a certain outcome being *x*, our credence that that outcome obtains should be *x*, absent more direct evidence bearing on the outcome.[1]

If, for example, I happen know that you are 90% reliable at arithmetic in the present circumstances, I should have 90% credence that you will answer an arithmetical question correctly.

We might then say that you and I are epistemic peers on a certain occasion with respect to a certain question just in case we are *equally reliable* at answering the question in the circumstances. The trouble is that while this may be a sufficient condition for epistemic peerage, we are seldom in a position to know that anything like it obtains. Much of the disagreement literature speaks of epistemic peers as those who are equals with respect to intelligence, thoroughness, and so forth. But this is not the ordinary case that is of most interest. Even in a case like arithmetic, where proficiency can be roughly measured (by test scores, say) I don't really know just how reliable you are. I can only give a rough estimate of how reliable I am, for that matter. You could easily be smarter than me. Then again, you might be dumber. You may have paid especially close attention to this particular question. Or perhaps you were uncharacteristically careless. None of this uncertainty of course gives me any reason to trust my judgment over yours in advance. For I have no more reason to suppose that I am more reliable than you than that you are more reliable than me.

We can model the situation better as seeing my credence as distributed over a range of possible degrees of reliability. Let's say that my reliability R_M can take any value in the set $\{0, .1, .2, .3, \ldots, 1\}$.[2] The *expectation* of my reliability is calculated as follows:

$$\mathrm{EXP}(R_M) = \sum_i i \ P(R_M = i)$$

$$(\text{Yours is } \mathrm{EXP}(R_Y) = \sum_i i \ P(R_Y = i)$$

The probability function P in each case here is the rational credence, or evidential probability for *me*, given my evidence. I will refer to these quantities as *my expected reliability* and *your expected reliability*. Keep in mind that the latter is the expectation of your reliability calculated in terms of *my* rational credence. (For the purposes of discussion, it's all about me, how I should respond to information that you disagree with me. You can switch it around for yourself.)

The shift from known reliability to expected reliability needn't make a big difference, for the following is a consequence of Reliability-Credence above:

> **Expected Reliability-Credence**: Absent more direct evidence bearing on the outcome, one's credence that a process will produce a certain outcome should equal the expected reliability of the process yielding that outcome.

To take a simple contrived case, my thermometer has two settings: 90% reliability or 70% reliability (don't ask me why). I'm 80% confident that I put it in the higher reliability setting. The expected reliability of the thermometer is $0.9 \times 0.8 + 0.7 \times 0.2 = 0.86$. This is how confident I should be in advance that the thermometer will give an accurate reading.

Thinking in terms of expected reliability fits well with Adam Elga's (2007) way of setting up the problem of disagreement, which I think is a useful way. As Elga uses the terms, I *count you as my peer* on a matter just in case my prior credence (prior to the details of the dispute) that my answer will be correct conditional on us

disagreeing on the answer is 1/2. It is worth noting that I can quite reasonably "count you as my peer" with respect to whether *p* in Elga's sense even if I am quite sure that you are *not* my peer as most philosophers following Kelly (2005) have understood it (where peerage involves equal reliability, or at least equal epistemic virtue). Suppose the teacher tells us that one of us is significantly better at arithmetic than the other, but not wanting any hurt feelings, she will not tell us who is better. I haven't seen the test scores so I have no idea if it is you or me. I'm sure we are not even close to being peers. But prior to doing a calculation, I have no reason to suppose that I will get it right rather than you in the event that we disagree. My prior credence that I will be right conditional on us disagreeing is 1/2. So I count you as my peer in Elga's sense. While I find Elga's approach useful, I believe the terminology has often lead readers astray ("count you as my peer" is naturally read as "believe you to be my peer"). But I don't know of a better way to put things, so I just ask that you keep it in mind that I'm following Elga's usage of "count you as my peer".

Here is a result linking expected reliability and peerage:

> **Expected Reliability-Peerage**: $EXP(R_M) = EXP(R_Y) \rightarrow P(\text{My answer is correct} \mid \text{We disagree}) = 1/2$.

Or in other words,

> **Expected Reliability-Peerage**: If you and I have equal expected reliability, then I should count you as my peer.

Perhaps that seems obvious enough already, but let me sketch a quick proof. Suppose you and I have the same expected reliability. There are just two ways for us to disagree: either I'm right and you're wrong, or you're right and I'm wrong. According to Expected Reliability-Credence, the prior probability that I/you will be right is equal to our expected reliability, let it be *r*. These probabilities are independent (you are no more or less likely to get it right given that I do). So P(I'm right & you're wrong) = P(You're right & I'm wrong) = $r(1-r)$. But now,

$$\begin{aligned} &P(\text{I'm right} \mid \text{We disagree}) \\ &= \frac{P(\text{I'm right \& you're wrong})}{P(\text{I'm right \& you're wrong}) + P(\text{You're right \& I'm wrong})} \\ &= 1/2. \end{aligned}$$

3. THE THERMOMETER MODEL AND EQUAL WEIGHT VIEW

Treating us both as thermometers seems to lead us to the following thesis.

> **Equal Weight View:** If prior to a dispute I count you as my peer, then upon learning that you have reached a different conclusion, I should be no more confident that I'm right than that you are.[3]

Combining this with points above, if you and I have the same expected reliability, then I can put no more confidence in my own conclusion than in yours. This seems hard to deny as long as I'm thinking of my own faculties as an indicator for me on a par with yours. If our two thermometers give different readings and I have no better grounds prior to this to expect mine to be right than yours, on what basis can I give preference to mine? Unless I have some further evidence besides the two thermometer readings, it would be foolish to give more credence to the reading of one thermometer just because it happens to be mine. Is the case of agents importantly different? You and I are given evidence *e* and asked to determine whether *p*. Antecedently I have no more confidence in your ability to answer such a question than mine. I conclude that *p*, you conclude that ∼*p*. The symmetry here seems to require the same response. To trust my convictions over yours seems as arbitrary as my trusting my thermometer just because it is mine.

There is reason to be suspicious of this way of viewing the matter. For notice that the initial evidence *e* on which my conclusion that *p* was based has dropped out of the picture. It may well be that the evidence supports my conclusion over yours. Of course the fact that you believe ∼*p* is (perhaps misleading) evidence that *e* supports ∼*p*. For insofar as I take you to be a reliable believer on the matter, I take you to be a reliable judge of the probative force of evidence (since like me you are in the habit of proportioning your belief to what you take to be the force of the evidence). But is this sufficient to neutralize, as it were, the force of the evidence, which may in fact favor my position over yours?

The situation might be dramatized in the following way.

> **The Oracle**: Having no idea whether *p*, I go to the Epistemology Oracle for help. The Epistemology Oracle doesn't always tell you what you want to hear, but always speaks the truth. She tells me, "Just two equally reliable agents inquire into the matter. One concludes that *p*, the other ∼*p*." That's not much help. Clearly I have no grounds to hold one conclusion over the other, having no reason to think either agent has a better chance of being right. Fortunately the Oracle adds, "The evidence on which they each based their conclusions is *e*."

Now this can make quite a difference. Evidence *e*, let's suppose, does in fact support *p*. There is no longer a straightforward evidential symmetry between *p* and ∼*p*. Perhaps the fact that there is disagreement on the issue should temper my confidence that *p*. But my total evidence counts in favor of *p*, so it would be unreasonable to remain neutral on the matter. But there's a twist.

> The oracle continues, "One of the epistemic peers that I mentioned is *you*. I was looking into the future to see that you would conclude that *p* on the basis of *e*."

According to the Equal Weight View, at this point the only rational response is to abandon my conviction that *p* entirely and become strictly agnostic on the matter. I must, in effect, ignore the evidence *e*, or "bracket" it off, not let it affect my judgment as to whether *p*. A moment ago I quite reasonably was more

confident that p than $\sim p$ on the basis of evidence e (even though the fact that there was disagreement on the matter was grounds for caution). Merely as a result of discovering the identity of one of the disagreeing agents (me), I can no longer rationally hold my view, despite the fact that I still have evidence e, which supports my conclusion. This can seem odd. Indeed, it can seem like irrational violation of the principle that one's convictions should always be based on one's *total* evidence.

Of course, the matter is not at all straightforward and I don't expect the story above to convince someone with a strong commitment to the Equal Weight View. They can at least offer the following explanation of why the identities of the disagreeing agents matter.

> Initially I assumed that the Oracle was referring to two people apart from me. Once I formed an opinion on the matter, I took it to be two against one, and two heads are better (more likely to be right) than one. The Oracle revealed that it was really one against one, a perfect tie between epistemic peers.

How satisfying is this response? It offers an explanation of the relevance of whether the disagreeing parties include oneself that fits with the thermometer model. But it does not address the concern that the Equal Weight View would have us ignore relevant evidence.[4]

4. HELP FROM BAYES?

Perhaps we can make progress on the matter by taking a Bayesian approach. In addition to its intrinsic loveliness, Bayesianism has proved fruitful in modeling the rational response to testimony more generally. We'll stick to the simple case in which we know that you and I have the same expected reliability of r at answering questions such as whether p, and we can expect that each of us will form a firm opinion on the matter one way or the other when we consider the same evidence e individually.

It might be tempting to think that there is a very short Bayesian step from our equal expected reliability to remaining agnostic in the event of a disagreement with you. As we saw above, given our equal expected reliability, my credence function should be such that $P(\text{I'm right} \mid \text{we disagree}) = 1/2$ (i.e. I count you as my peer). Now as I examine evidence e and conclude that p, while you conclude that $\sim p$, I discover that we disagree. Shouldn't I now just conditionalize on this information and set my new credence that I'm right to my prior conditional credence of $1/2$?

This is misleading. Properly understood, a rule of conditionalization states that I should conditionalize on the *strongest* proposition that I learn. Suppose I know that yesterday you and I were given the same evidence and asked whether p, but with partial amnesia I have no recollection of what the evidence was, what either of us concluded, or any other details (although I still count you as my peer). Given that I take myself to be quite reliable on the matter, I figure that whatever I answered was probably correct and that most likely you agreed. But I ask the examiner just the following: "Did we agree or disagree?" "You disagreed," he replies. Now this

does seem to be a straightforward matter of conditionalization. Learning only we disagreed, my confidence that my answer (whatever it was) is correct should drop to my prior conditional credence of 1/2. But now the examiner goes on, "The evidence you were given was *e*, and the conclusion that you drew was that *p*." And now it all comes back to me, how *e* supports *p* (which in fact it does, let's suppose). The controversial question is how, if at all, any of this further information should affect my opinion. This is not settled by the fact that I had prior conditional credence P(I'm right | we disagree) = 1/2, and a modest rule of conditionalization.

A natural way to approach the matter is to suppose that having first come to my own opinion that *p*, I conditionalize on the information that *you believe* ~*p*. That is the total relevant information that I learn, and this is the standard way that Bayesians have approached the more general topic of learning from testimony. Given that your expected reliability on the matter is *r*, P(you believe ~*p* | ~*p* & *e*) = *r*, and P(you believe ~*p* | *p* & *e*) = P(~you believe *p* | *p* & *e*) = 1−*r* (since you will either conclude that *p* or conclude that ~*p*). Now,

$$
\begin{aligned}
&\mathrm{P}(p \mid \text{you believe } {\sim}p \,\&\, e) \\
&= \frac{\mathrm{P}(p \mid e)\,\mathrm{P}(\text{you believe } {\sim}p \mid p \,\&\, e)}{[\mathrm{P}(p \mid e)\,\mathrm{P}(\text{you believe } {\sim}p \mid p \,\&\, e) + \mathrm{P}({\sim}p \mid e)\,\mathrm{P}(\text{you believe } {\sim}p \mid {\sim}p \,\&\, e)]} \\
&= \frac{\mathrm{P}(p \mid e)(1 - r)}{\mathrm{P}(p \mid e)(1 - r) + \mathrm{P}({\sim}p \mid e)r} \\
&= 1/2 \quad \text{iff} \quad \mathrm{P}(p \mid e) = r.
\end{aligned}
$$

If I'm updating my attitude to *p* by conditionalizing on the information that you believe otherwise, then I will give no more credence to my view than to yours only if my credence in *p* given evidence *e* prior to learning of your contrary opinion was equal to my expected reliability. On a Bayesian approach then, the Equal Weight View is committed to the following constraint on my own credence prior to disagreement.

> **Calibration Rule:** If I draw the conclusion that *p* on the basis of any evidence *e*, my credence in *p* should equal my prior expected reliability with respect to *p*.

This can seem like a bit of common sense. I have a long track record of getting about 90% of my 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%.

So the Calibration Rule has its appeal. But there is cause for concern along now familiar lines: evidence *e* seems to drop out of the picture, as the rule puts a constraint on my credence for any evidence I might have. Now if the evidence is not to play its usual role in governing how I proportion my belief, one might wonder why I should bother considering the evidence at all if at the end of the day the Calibration Rule just has me set my credence by a factor (expected reliability) that was available before I even obtained the evidence. But this is not quite a fair way to put it. The Calibration Rule has my final credence in *p* be a function not only of my prior expected reliability but also of *my initial judgment on the matter*. If my expected reliability with respect to *p* is *r*, then by the Calibration Rule my credence in *p* must end up as either *r* or $(1-r)$. Which one it will be depends on whether the belief that I form is that *p* (in which case my credence that *p* shall be *r*) or $\sim p$ (giving me a final credence of $(1-r)$ that *p*. Now when it comes to forming this initial judgment, no doubt a proponent of the Calibration Rule will encourage us to do so on the basis of all the evidence available. After all, basing my conclusion on the evidence surely improves my chances of getting the answer right. And the expectation of my reliability that is relevant here is my reliability given that I'm doing my best to get the answer right (rather than pulling my conclusion out of a hat, or out of somewhere else).

If all goes well then, in following the Calibration Rule I will consider the evidence *e* and form an opinion on whether *p* that is in fact warranted by the evidence. Suppose that *e* strongly supports *p*. But my expected reliability for *p* is only 70%. Ideally, when I examine the evidence, I will proportion my belief to the evidence and become very confident that *p*. But then, in the light of my expected reliability, I should rein my credence in *p* back down to 0.7. In the best case, the evidence is only partially determining my attitude to *p*. What we might call the *direction* of the evidence – whether it "points" in the direction of *p* or toward $\sim p$ – determines whether my credence will be 0.7 or 0.3. The *strength* of the evidence – in this case, the fact that it strongly supports *p* – has no role to play in determining my attitude. But now if the evidence *e* is to play *some* role in determining my credence, why shouldn't it just do the whole job? And if the evidence has to divide up its belief-determining duties with my expected reliability, why is the evidence relegated to this task (of selecting between credence *r* and $1-r$) in particular?

All very mysterious. But our task for now is still to see how the Calibration Rule fares in the thermometer model on a Bayesian approach. We have been supposing that credence in the light of disagreement with you should be the result

of conditionalizing on the information that *you believe* $\sim p$, just as I might update my opinion as I read your thermometer. Now if we take the thermometer model seriously, I should apparently set my credence in p prior to learning of your opinion in a similar way: by conditionalizing on the information that *I believe that p*. That's what I do with thermometers. I form my opinion on the temperature by first updating on the thermometer reading I obtain and the updating further with the reading from yours (of course the order doesn't really matter).

The expected reliability of my thermometer is 90% say. So naturally I'm 90% confident that the reading it gives me will be correct. Now if it gives me a reading of 60°, should my credence that it is 60° out be 0.9, as an analogue of the Calibration Rule would seem to suggest? Well, no, not exactly. Typically my confidence in the thermometer's reading should be a function not only of its expected reliability but will also depend on what the reading is. Bayesian updating always proceeds from some prior probability assignment. Let's imagine an even simpler thermometer model. My thermometer has only two possible readings: HOT and ~HOT ("HOT" is just short for "more than 90° F").

Let h = the thermometer reads HOT, e be my total evidence prior to reading it, and r be the expected reliability of the thermometer.

$$
\begin{aligned}
&\mathrm{P}(\textit{It's hot} \mid h \,\&\, e) \\
&= \frac{\mathrm{P}(\textit{It's hot} \mid e)\,\mathrm{P}(h \mid \textit{It's hot} \,\&\, e)}{\mathrm{P}(\textit{It's hot} \mid e)\,\mathrm{P}(h \mid \textit{It's hot} \,\&\, e) + \mathrm{P}(\sim\textit{It's hot} \mid e)\,\mathrm{P}(h \mid \sim\textit{It's hot} \,\&\, e)} \\
&= \frac{\mathrm{P}(\textit{It's hot} \mid e)r}{\mathrm{P}(\textit{It's hot} \mid e)r + \mathrm{P}(\sim\textit{It's hot} \mid e)(1 - r)} \\
&= r \quad \text{iff} \quad \mathrm{P}(\textit{It's hot} \mid e) = 1/2.
\end{aligned}
$$

If my prior credence that it is hot out is 1/2, then when I conditionalize on the thermometer's reading of HOT, my credence that it is hot should be updated to equal its expected reliability. But this is a special case. Usually, I will be more confident than not that it is hot or that it is not before I look at the thermometer. Even if I know the thermometer is highly reliable, say 99%, if I'm fairly confident that it is cold before reading the thermometer, then the fact that the thermometer reads HOT is some evidence that it is not accurate on this occasion and my subsequent credence that it is hot should be less than 99%.

Can this kind of picture be applied to thinking agents? Suppose that having concluded that p on the basis of evidence e, I'm to update on the information that *I now believe that p* (where my expected reliability is r). The same result holds:

$$\mathrm{P}(p|e \,\&\, \text{I believe}\, p) = r \quad \text{iff} \quad \mathrm{P}(p|e) = 1/2.$$

So on this Bayesian model the Calibration Rule holds only if my credence in p, given the evidence but prior to updating on the information that I believe that p,

should be 1/2. But this is incoherent. Once I've formed a *belief* as to whether *p*, I can hardly have *credence* 1/2 that *p*! If I do update from whatever credence I have when I conclude that *p* (which may be quite high), then I will end up with more than *r* confidence in *p*. (And when I further update on the fact that you believe ~*p*, I will put more confidence in my answer than yours). Perhaps a friend of the Calibration Rule who wants it to fit in with a Bayesian approach will want to say that the prior credence that one must update from, given the information that I now believe that *p*, is one prior to my having formed a belief on the matter. Perhaps my credence in *p* before I even obtained any evidence should have been 1/2. We could think of it as follows. Suppose I know that yesterday I formed a belief as to whether *p*, but I've forgotten what I concluded and why. Without any evidence to go on, my credence in *p* is 1/2. I'm told that yesterday I concluded that *p*. Given the result above, when I update on the fact that I concluded that *p*, my credence in *p* will shift to equal my expected reliability, just as the Calibration Rule recommends. It might be thought that in the typical case I should update in a similar way.

This story involves quite a departure from standard Bayesian updating, as I'm updating my credence in *p* only on the fact *that I believe that p*, and not the *evidence e* on which I first based my belief. So it can hardly be considered as a Bayesian *defense* of the Calibration Rule. There is a further problem. This account depends on the thesis that prior to the evidence, my credence in *p* should be 1/2. It is tempting to suppose that without evidence to go on, my attitude should be neutral with respect to *p* and hence 1/2 is the appropriate credence. I'm not unsympathetic to the thought here (see White forthcoming). But it is not an option on standard probability to assign 1/2 prior probability to *all* propositions.

It might be suggested that while responding to the information that *you* my peer believe ~*p* is best modeled as a matter of Bayesian updating, the information that *I* believe *p* should not be handled in the same way. The Calibration Rule might be thought of as a fundamental rule of rational credence that cannot be understood as a special instance of conditionalization. The trouble is that to take this line requires abandoning the initial motivation for the Equal Weight View. According to the thermometer model, I am to treat myself as a thermometer just as I treat you. In treating my rational faculties as an indicating device on a par with yours, I am to treat the information about what my belief forming mechanisms produced in the same way that I treat yours. To insist that the Calibration Rule is a fundamental extra-Bayesian constraint is to suppose that when it comes to information about what people believe, the first-person and third-person cases are importantly different.

5. BOOTSTRAPPING

Aside from commitment to the thermometer model, can anything be said for the Calibration Rule? Here is an argument in the same spirit as Elga's (2007) bootstrapping argument for the Equal Weight View, only applied to the case of

the single agent. (It is an elaboration of the initial motivation for the Calibration Rule that I sketched above.)

> **Arithmetical Bootstrapping:** My expected reliability for a series of True/False arithmetic questions is 90%. The Calibration Rule would have me put 0.9 confidence in each of the answers that I give. I violate the rule and put 0.95 credence in each answer.

Now if this 95% confidence in my answers is acceptable, it seems appropriate to increase my expectation of my reliability above 90%. Given the large number of questions, if I'm 95% confident of each answer, then I should be very confident that *about* 95% of my answers are correct. But then it is far more likely that about 95% of my answers are correct if I'm about 95% reliable than if I'm only 90% reliable. So I should shift my credence more toward those hypotheses that assign me a higher reliability and hence my expected reliability increases above 90%. I can now be more than 90% confident that I will answer correctly the next time. I have increased my expectation of the reliability of my reasoning faculties and my confidence that I will answer correctly in the future, just by the use of that very faculty. This seems dubious, to say the least. To do this with a thermometer would seem mad. Surely, to increase the expected reliability of my thermometer, I need some *independent* check on the correctness of its answers.

Can we tolerate the conclusion that I can rationally increase my confidence in my arithmetical abilities just by doing some arithmetic and judging for myself whether those answers are correct? Following others, I've used the possibility of bootstrapping as an objection to views in the epistemology of perception that allow one to gain perceptual justification without independent justification for one's own perceptual reliability.[5] According to a simple version of Dogmatism (Pryor 2000), provided I have no reason to suspect that my color vision is deceiving me, a visual experience as of a red card is sufficient to justify me in believing that it is red, even if I lack antecedent justification to suppose that my color vision is reliable.

> **Color Bootstrapping:** I view a series of colored cards, judge them to be the color they appear to be (this one's red and it appears red, this one is blue and it appears blue, …), and hence conclude that my color vision is likely to be reliable (since it keeps on correctly representing the color of the cards).

This seems obviously absurd, and bootstrapping in the arithmetic case can seem just as bad. Still, it is worth noting two respects in which the arithmetical bootstrapping isn't *so* obviously bad. First, if I accept Dogmatism then I should predict in advance of even looking at the cards that I will gain justification to suppose that my color vision is reliable when I view the cards regardless of how they appear. But this cannot be right. A course of experience cannot justify a conclusion if I would be so justified regardless of what experience I have. If I can tell in advance that I will end up being justified in believing that my color vision is reliable, then actually seeing the cards is redundant. I might as well just go ahead

and believe that my color vision is reliable without bothering to investigate at all. But that is absurd.

Now denying the Calibration Rule needn't have this consequence. On the most plausible alternative to this rule, it is appropriate for me to give higher than .90 credence to my answer *only if my evidence does in fact support my conclusion*. On this view, all I know in advance is that *if* I correctly give the answers that are in fact supported by the evidence, then I will be justified in increasing my expectation of my reliability at arithmetic. But there is no guarantee that I will do so. Indeed, I should only expect to do so about 90% of the time. So perhaps that's not quite as bad.

Second, consider a slightly different way of bringing out what seems absurd about the color vision bootstrapping. Suppose an Oracle tells you in advance that the first card will appear *red* to you. (The Oracle can just see into the future and directly tell with certainty how your experiences will be. She doesn't have to base her prediction on her knowledge of what color the card actually is). No one would suppose that this information has any bearing on how reliable your color vision is. You knew it was going to appear some color or other. There is nothing about red-appearances that tells you anything about the reliability of your color vision. Now you take a look at the card and, sure enough, it appears red as you were told. Have you *now* learnt anything relevant to your reliability? There are of course some important differences between your *knowing that the card will appear red* and its *currently appearing red to you*. But it would be strange to respond,

> I already knew that this card would appear red, as it now does. And that by itself was a little evidence that it is red (since I knew my color vision was likely to be at least *somewhat* reliable). But now that I look at the card, I see that it is indeed red! So now I'm more confident that my color vision worked properly this time.

It seems clear that once I know that the card will appear red to me, I gain no further evidence regarding its color (let alone about the accuracy of my color vision on this occasion or its general reliability) upon its so appearing to me. Since the Oracle's report that the card will appear red to me is irrelevant to how reliable my color vision is, the bootstrapping procedure should have no effect on my expected reliability.

Let's set the case up similarly with respect to the Arithmetic case. The Oracle reveals that when I examine the evidence e for the first question, it will *seem* to me (rightly or wrongly) that the answer is p. Having no clue whether p in advance of examining the evidence, I can hardly take the Oracle's report as evidence of my reliability at arithmetic. That I will come to believe that p is some evidence for me that p. But my confidence in p at this point must be calibrated with my rational expectation of my reliability on the matter.[6] Now I go ahead and do the calculation, arriving as expected at the answer p.[7] Could it be reasonable to respond as follows?

> I already knew that the answer to this question would *seem* to me to be *p*, as it now does. And that by itself was a little evidence that the answer is *p* (since I took my arithmetical abilities to be fairly reliable). But now having done the calculation, I see that the answer is indeed *p*! So now I'm more confident that my calculation worked properly this time.

Such a response might seem just as bad as in the color-vision case. But there is an important difference between the cases to note. In the color vision case it is very implausible to claim that I gain any new relevant evidence upon seeing the red card. My evidence just seems to consist in facts about how the card appears to me. And I've already been given this in advance by the Oracle. In the arithmetic case my evidence need not consist merely in facts about which answer *seems correct* (this much is given in advance by the Oracle). But it should be uncontroversial that as I do the calculation, I do obtain *additional relevant evidence* bearing on whether *p*. When I answer the arithmetic question, I am given evidence *e* which may in fact be decisive evidence for *p*. Of course you may be suspicious that while I do obtain this evidence, it should not be pushing my confidence in my answer above my expected reliability. This is the tricky question that we keep coming back to. But my point here is just that a bootstrapper who chooses to bite the bullet in the color-vision case and say that he is entitled to boost his confidence that the card is red when he sees it faces the challenging question: "What evidence have you gained that would justify such a boost?" The bootstrapper who violates the Calibration Rule in the arithmetic case has an easy answer, "*e*".

6. How Conciliatory is the Thermometer Model?

The thermometer model would appear to have radically conciliatory consequences. If I count you as my peer and I am to treat myself as an indicator on a par with you, it appears that I can put no more confidence in my own conclusions than in yours. Furthermore, it appears that I can't *demote* you from the status of peer. That your thermometer gives a different reading than mine may be some evidence against its reliability. But surely it is equally evidence against the reliability of mine. I want to consider more carefully whether this is right and see to what extent we can resist a radically conciliatory position on the thermometer model.

You and I have binary thermometers that read either HOT or ~HOT when placed in a beaker of liquid. Let's also suppose that half of the samples in the beakers are hot and half cold. So apart from the readings we get from one or both thermometers, we have no evidence either way as to whether a sample is hot or cold. We can suppose that prior to reading the thermometer, my credence that the sample is hot is 1/2. The lab assistant first tells me whether the thermometers agree or disagree. Given the expected reliability of the two thermometers, I can calculate the probability that my thermometer gave an accurate reading. I might further learn that my thermometer read HOT. Since the prior probability that the sample is hot

is 1/2, my credence that it is hot given that thermometer reads HOT and yours reads ~HOT should just equal my credence that my thermometer's reading is *correct* given just that our thermometers disagree. That is,

$$\begin{aligned} &P(\text{Sample is hot} \mid \text{my thermometer reads HOT \& yours reads } \sim\text{HOT}) \\ = {} &P(\text{my thermometer reads correctly} \mid \text{our thermometers disagree}) \end{aligned}$$

(If it was a sample that was initially more likely to be cold than hot, that my thermometer reads HOT would be some evidence that its reading is not correct on this occasion, and this equality would not hold). Given this unproblematic step from the probability of an accurate reading to the probability of the temperature of a sample, we can simplify our discussion by focusing just on questions of the accuracy of a reading given information about the agreement or disagreement of the readings.

Now, as we saw back in Section 2, if the expected reliabilities of our two thermometers are the same, then conditional on them disagreeing the probability that mine is correct must be 1/2. This is a strongly conciliatory result. Perhaps I know that mine is very reliable, and hence am very confident in the reading it gives me when I read it alone. But once I see that yours disagrees, my attitude should revert to agnosticism. Here is some comfort for those of us who don't like being so spineless: It is easy to underestimate the impact that small asymmetries in expected reliability can have on the result. To take a toy case, suppose we know that my thermometer is 99% reliable whereas yours is 95% reliable (perhaps having been bumped around is enough to reduce its reliability just a little). Since both thermometers are highly reliable, it is tempting to think that we should put just a little more confidence in my thermometer in the event of a disagreement between them. After all, your thermometer is almost always correct and given its reading alone, we ought to be very confident in the answer it gives (only slightly less confident than if we had only read my thermometer). So it might seem arbitrary and unreasonable to drastically dismiss your thermometer reading when it disagrees with mine.

But this is a mistake. Given the disagreement, the probability that my thermometer is correct is as follows:

Letting T_i = Thermometer *i*'s reading is correct

$$\begin{aligned} P(T1 \mid \text{thermometers disagree}) &= P(T1 \,\&\sim T2)/[P(T1 \,\&\sim T2) + P(\sim T1 \,\& T2)] \\ &= (.99 \times .05)/(.99 \times .05 + .01 \times .95) \\ &\approx .84 \end{aligned}$$

If I read my thermometer first, then upon reading yours my confidence in my reading should drop somewhat (from .99 to .84). But I should remain fairly confident that my thermometer was right. While I began being highly confident

(95%) that your thermometer would give an accurate reading, once I see how it disagrees with my own, I should drastically reduce my confidence, indeed becoming fairly sure that your thermometer is mistaken on this occasion.

Cases involving thinking agents rather than thermometers are of course a little more messy. But similar points apply. First, let's keep in mind that reliability is not just a matter of how accurate my answers tend to be on average over time, but how disposed I am to answer correctly given the circumstances of this occasion. Second, it is expected reliabilities that matter here, not known reliabilities. Now, it doesn't take much to have a slightly higher expectation of reliability for myself than for you. Perhaps you and I have equal (and very good) track records at basic arithmetic. There is no reason to suppose that in general I will get arithmetical questions right more often than you. But as others have pointed out, on a particular occasion I may have information about myself that I lack about you.[8] I might be aware that I'm alert, have paid careful attention to the question, and so forth. While I have no particular reason to suppose that you might reason poorly on this occasion, I do not have the same grounds for confidence in you as I have in myself. This can be enough to justify a slight difference in expectations of reliability. And given the kind of result sketched above, this can be enough to warrant significantly more confidence in my answer over yours if we disagree.

Unfortunately, in many cases I am not warranted in having a higher expectation of my reliability than yours. Here we have symmetry of expected reliability, so there appears to be no room for me to favor my own opinion. However, there may still be a different kind of asymmetry, an asymmetry of *resilience*, which I will try to illustrate by returning to the toy thermometer model. My thermometer is known to be 90% reliable. Yours has three reliability settings: 85%, 90%, and 95% reliability. I have no idea which setting your thermometer is on, and my credence is 1/3 that it is on any particular setting. While I know more about my thermometer than yours, the expected reliability of each thermometer is 90%. If I were to pick one thermometer to use I should have no reason to prefer mine over yours. But now suppose we each use our own thermometer and get different readings. Since I *know* that mine is 90% reliable, this information should have no effect on its expected reliability, which remains at 90%. The fact that we disagree is, however, relevant to the reliability of your thermometer. The more reliable the two thermometers are, the more likely they are to each give the correct reading and hence to agree. A disagreement between the two readings is more to be expected if our thermometers are 90% and 85% reliable than if they are both 90%, or 90% and 95% reliable. Hence the information that the readings disagree should shift my credence more towards your thermometer being 85% reliable than 90% or 95%. And so my expectation of your thermometer's reliability should be lowered relative to mine. In this way I can legitimately demote you in my judgment from the status of peer (i.e. I will no longer count you as my peer in Elga's sense) without antecedently having any more reason to suppose that you would be mistaken than that I would.

How might we apply the toy model to thinking agents? In a typical case my credence should be spread over a range of possible degrees of reliability both for me and for you. And these may average out to the same (i.e. have the same expectation value). Nevertheless, my credence distribution for my own degree of reliability may be more concentrated around a certain value, while my distribution for your reliability may be spread more thinly. Let's say I have some idea of how reliably I can judge whether p and my best guess is about 90%. For all I can tell, it might be *somewhat* higher or lower. But as we go further from the 90% mark, it becomes significantly less likely that that is how reliable I am. I can be pretty sure that I'm neither a genius nor an idiot. But I can't be quite so sure about you. I'm not in your shoes and know less about you than I do about myself. It might be that you are enjoying an episode of exceptional mental clarity and insight, or it might be that you are temporarily deranged or high. While neither is particularly likely, it is even less likely that I am in either extreme state, given what I can judge by introspection. Similarly, it could be that you have some decisive evidence for your position that I'm not aware of. Or perhaps you are taking a wild guess on very little evidence. (The more we have discussed the matter, the more likely it is that we have the same or similar evidence. But there is no guarantee of this.) Again, while this may be unlikely, I can be even more confident that my evidential situation is not at these extremes, given what I know about myself. In general, it seems that as a result of having better familiarity with my own situation than with yours, my credence distribution will tend to be "spiked" more around a certain degree of my own reliability, and more flat with respect to yours. The result is that my expectation of my own reliability will be more *resilient* than my expectation of yours, in the sense that it will shift less dramatically in response to evidence against each of our reliabilities. The fact that we have given different answers to a question counts as evidence against our reliabilities. For the more reliable each of us is, the more likely we are to agree on the truth. Given this difference in resilience as described, my expected reliability should be reduced somewhat, but less dramatically than your expected reliability should be reduced. I should end up with a greater expectation of my own reliability than of yours.

This appears to be a rather comforting result. Perhaps a bit too good to be true. I began with no more confidence that I would answer correctly than that you would. But now, upon learning that we disagree, I have greater expectation of my own reliability than of yours (albeit less than my prior expectation of my own reliability). It appears as though I am now entitled to put more confidence in my own *answer* than in yours. Having demoted you with respect to reliability at answering correctly, I should now judge you less likely to be correct. Right?

Wrong. Perhaps surprisingly, while the information that we disagree should lead me to lower my expectation of your reliability relative to mine, this does not warrant giving more credence to my being right than you. Going back to the thermometer model, as we have seen, the calculation of the probability that my thermometer's reading is accurate given that the readings disagree is straightforward. There are two

ways in which the thermometers can disagree: mine is right and yours is wrong, or yours is right and mine is wrong. The prior probability of the correctness of either thermometer reading is just 0.9, and these probabilities are independent. So the prior probability of either disagreement outcome is $P(T1 \ \& \sim T2) = P(\sim T1 \ \& \ T2) = 0.9 \times 0.1 = .09$. Neither outcome is more probable conditional on one of them obtaining.

So while the asymmetry of resilience can lead me to demote you from peerhood status, it does not allow me to put any more confidence in my own opinions. But what if our disagreements continue? Suppose, having obtained different readings from our thermometers, we try them again on another sample and find different readings again. Given the greater expected reliability of my thermometer, this time around I should be more confident that my answer is correct. Indeed, given the two disagreements, it is somewhat more likely that my thermometer was correct on the first reading also. In this way, if we continue to disagree, I can increasingly put more confidence in my own answers and judge you to be less reliable.[9] Perhaps this still seems a bit too good to be true. It appears that if we agree I can trust our shared opinion, and if we disagree I can still favor my own opinion over yours. But there is a flip side to it. If our thermometers *agree* on a reading, this is evidence that your thermometer is more than 90% reliable. In response to the information that our thermometers agree, I should *increase* my expectation of your thermometer's reliability relative to mine. So then if we get conflicting readings on other occasions, I should give more credence to your answers than to mine. Perhaps being thermometers isn't so bad.[10]

REFERENCES

Christensen, David. 2007. "Epistemology of Disagreement: The Good News." *Philosophical Review* 116: 187–217.

Christensen, David. Forthcoming. "Higher-Order Evidence." *Philosophy and Phenomenological Research.*

Christensen, David. Manuscript. "Disagreement, Question-begging, and Epistemic Self-criticism."

Cohen, Stewart. 2002. "Basic Knowledge and the Problem of Easy Knowledge." *Philosophy and Phenomenological Research* 65: 309–29.

Elga, Adam. 2007. "Reflection and Disagreement." *Nous* 41: 487–502.

Kelly, Thomas. 2005. "The Epistemic Significance of Disagreement." In T. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology*, vol. 1, pp. 167–96. Oxford: Oxford University Press.

Kelly, Thomas. Forthcoming. "Peer Disagreement and Higher-Order Evidence." In R. Feldman and T. Warfield (eds.), *Disagreement.* Oxford: Oxford University Press.

Lackey, Jennifer. Forthcoming. "A Justificationist View of Disagreement's Evidential Significance." In A. Haddock, A. Millar, and D. Pritchard (eds.), *Social Epistemology.* Oxford: Oxford University Press.

Lewis, David. 1980. "A Subjectivist's Guide to Objective Chance." In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, vol. 2. Berkeley: University of California Press.

Pryor, James. 2000. "The Skeptic and the Dogmatist." *Nous* 34: 517–49.

White, Roger. 2006. "Problems for Dogmatism." *Philosophical Studies* 131: 525–57.

White, Roger. Forthcoming. "Evidential Symmetry and Mushy Credence." In T. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology*, vol. 3. Oxford: Oxford University Press.

NOTES

1 This, of course, is just a crude statement of Lewis' Principal Principle (1980).

2 We could take my credence to be a continuous function over degrees of reliability, and the expectation as an integral. But there no need to get too technical.

3 Elga (2007) defends the Equal Weight View, but not in a way that leans on the thermometer model.

4 Kelly (forthcoming) develops an appeal to total evidence very effectively. For a subtle defense of the need to "bracket off" evidence we possess, see Christensen (forthcoming).

5 I develop bootstrapping objection to Dogmatism in White (2006), following Cohen (2002).

6 We can suppose that my prior credence in p in this case was 1/2. Then it should be straightforward that I should conditionalize on the Oracle's report and update my credence to my expected reliability of 0.9.

7 It must be strange to reason about a question while knowing all along what it is you are going to conclude. Perhaps this will affect your reliability. Let's avoid these interesting complications by supposing that I temporarily forget about the Oracle's report while I calculate.

8 See Christensen (2007) and Lackey (forthcoming).

9 Christensen (manuscript) develops a similar idea less formally.

10 Thanks to Nathan Barczi, David Christensen, Adam Elga, Tom Kelly, and Katia Vavova for discussions on this topic, and to audiences at ERG and the *Episteme* Conference at Northwestern University for feedback.

Roger White is an Associate Professor in the Department of Linguistics and Philosophy at MIT. His research interests include problems in probability and confirmation theory, the role of explanatory considerations in theory choice, and foundational questions concerning the nature of epistemology.