Final version published in *Episteme*, (2016): 397-422. Please feel free to email me for a pdf of the published version: david.christensen@brown.edu

Disagreement, Drugs, etc.: From Accuracy to Akrasia[1]

David Christensen

Introduction

We often get evidence concerning the reliability of our own thinking on some particular matter. This happens frequently when we learn about factors that can impair our cognitive reliability. Examples of such evidence are pretty common and obvious: think of what we know about the epistemic effects of insufficient sleep or oxygen, of alcohol and other drugs, of emotional ties to other people, or of implicit biases based on race, nationality, or gender. And spending a few hours reading a book like Kahneman (2011) reveals a depressing plethora of unexpected and unobvious factors that subtly but significantly disrupt reliable cognition. Once one gets evidence that such a factor may be influencing one's own belief on some matter, it seems that rationality requires one to take account of this evidence. In many cases, this will require adjusting one's credence.

I think that something similar is true about a kind of evidence that has attracted a lot of attention recently: evidence provided by the disagreement of others. The connection is particularly clear in cases where one has good reason to regard those who disagree as reliable thinkers who share one's evidence relevant to the disputed matter. The disagreement may not point to any particular physiological or psychological factor as likely to have caused one to think unreliably. But it makes it more likely that one's thinking has not been reliable, for some reason or other. And as in the case of particular disruptive factors, it seems that rationality can require one to modify one's credences in response to this sort of evidence.

Of course, these are not uncontroversial claims. Let us call this sort of evidence—roughly, evidence that constrains one's credence on some subject-matter via bearing on the reliability of one's thinking about that subject-matter—higher-order evidence. Some have doubted that rational belief is sensitive to higher-order evidence.[2] But I'm going to presuppose that rational belief is sensitive to higher-order evidence, and that it would be good to work out rational norms for responding to this

[1] Ancestors and earlier drafts of this paper were presented at the Center for Advanced Studies, Ludwig Maximilians-Universität; the Midwest Epistemology Workshop at the University of Michigan; Brandeis University; Northwestern University; the Episteme Conference in Phuket, Thailand; the University of Rochester; the University of Copenhagen; Simon Fraser University; Union College; the University of Vermont; and LOGOS at the University of Barcelona. I'd like to thank the audiences at those presentations, and my commentator at the Episteme conference, Lauren Leydon-Hardy, for valuable comments and questions. Special thanks to Dominik Berger, Stew Cohen, Jeremy Fantl, Sophie Horowitz, Zoë Johnson-King, Sarah Moss, Baron Reed, Blake Roeber, Andrew Rotondo, Josh Schechter, Miriam Schoenfield, Robert Steel, and Jonathan Vogel for very helpful discussions of these issues and/or comments on earlier drafts.

[2] For some examples of this position (or closely-related positions), see Kelly (2005), Lasonen-Aarnio (2014), Schoenfield (2014), or Titelbaum (2015).

kind of evidence. So here, I'll try to make progress on this front. I'll look at some plusses and minuses of two fairly general models for accommodating higher-order evidence. [3]

As it turns out, the model that currently seems most promising to me also has interesting implications for the question of whether it can be rational to be epistemically akratic. Epistemic akrasia might be glossed as believing something while thinking that it's not the epistemically rational thing for one to believe in one's situation. A little less roughly, we might gloss it as having a certain belief, or degree of belief, while also thinking that some particular other belief, or degree of belief, would be more epistemically rational in one's situation. It is natural to doubt that it can be rational to be epistemically akratic in roughly this way. After all, if I'm confident that there's some other belief that would be more rational in my situation, how can I rationally maintain my current belief? Wouldn't rationality require me to adopt the belief that I'm confident is more rational? Epistemic akrasia may even seem Moore-paradoxical: It's certainly not inconceivable that P is true even though my total evidence supports ~P. But there seems to be something wrong with me believing "P, but my total evidence supports ~P". Some (Adler 2002) have even argued that clear-eyed epistemic akrasia is simply impossible. However, it now seems to me that reflecting on higher-order evidence may show us that, in some strange situations, and even in a number of fairly ordinary situations, epistemic akrasia may be not only permissible, but rationally required.

I should note that higher-order evidence often comes mixed with ordinary evidence. In fact, I doubt that in such cases we can always even in principle divide the evidence up into the purely higher-order bits and the first-order bits. I suspect that the best way to think about things will turn out to be in terms of different ways that evidence bears on a proposition, rather than in terms of different bits of evidence. Here, I will sometimes be careful about this, but I'll mostly write as if the bits of evidence can be separated. While this no doubt can cover up some real difficulties, I hope that simplifying the discussion in this way will allow other difficulties to be studied more easily. But I do think that a fully satisfactory account of higher-order evidence will require better understanding of this issue.

1. Disagreement, and setting up the question

Let's begin with a disagreement example, which will set up the main question:

Logic Disagreement: Alison is told the following two things by an impeccable source:

(P1) Karla went to the party if and only if Kayla didn't; and

[3] I've tried in the past to say somewhat general things about how higher-order evidence should be accommodated (Christensen 2010, 2011), but I've never tried to develop really general principles. The present paper is heavily influenced by two papers which do deal directly with that issue. Paulina Sliwa and Sophie Horowitz (2015) discuss two general principles they call "Guess Calibration" and "Evidential Calibration," and endorse the latter. Miriam Schoenfield (2014) poses difficulties for several proposals, including Sliwa and Horowitz's. The principles I'll consider are importantly similar to the ones they discuss, though they're also importantly different. Related ideas are also discussed in White (2009). I won't have space here to discuss the similarities and differences properly, but I recommend these papers to interested readers.

(P2) Either Kayla went or Layla and Lola, the Lumpkin twins, both went.

Alison thinks about these facts, and becomes confident that

(C): If Layla Lumpkin didn't go, then neither did Karla.

Then Alison learns that her friend Ben, who has a long and excellent record of accuracy in exactly this sort of reasoning, has a low credence in C. But this is not because Ben doubts either of the premises Alison started with. In fact, Ben is highly confident of both, but doesn't think that they entail C. Alison reflects on the fact that she makes mistakes in this sort of reasoning at least as often as Ben does. This is true even when—as in the present case—it seems to Alison that she can clearly see why certain premises entail a particular conclusion. And so Alison becomes significantly less confident in C.

A common reaction to such cases is that it is rational for Alison to have a very high pre-disagreement credence in C, but after learning of Ben's disagreement, she must reduce her confidence considerably. Let us suppose that this reaction is correct. How might we explain it?

Part of what's relevant is some sort of assessment Alison should make about the likelihoods of Ben having made a mistake and herself having made a mistake. For example, if she has good evidence that Ben is generally a much more reliable logician than she is, Alison should assess him as less likely to have made a mistake, and should lose more confidence in C, than she should if she has evidence that he is generally equally reliable. And the information that bears on the question of who's likely to have made a mistake today need not be restricted to information about general reliability. For example, if Alison knows that Ben is drunk, she should think him more likely to have made a mistake than otherwise, and lose less confidence in C than otherwise.

That said, there seems to be an important limitation on the factors that can bear on the assessment it's rational for Alison to make of Ben's reliability. It would seem irrational for her to dismiss Ben's disagreement completely by reasoning as follows:

Quick Argument for dismissing Ben's disagreement:

1. P1
2. P2

So, 3. C

4. Ben believes ~C

So, 5. Ben is wrong this time!

What seems wrong here is that this would dismiss Ben's disagreement on the basis of the very reasoning whose correctness was brought into question by Ben's dissent. If rationality requires Alison to take the evidence of Ben's dissent seriously, what seems to be needed is an assessment of Ben's likelihood of error that's *independent* of that reasoning. The need for this sort of independence

is particularly clear in examples involving disagreements over logic. After all, the Quick Argument above has unquestioned premises and is clearly valid. So it's not the sort of argument whose force will be affected by adding premises that provide probabilistic empirical considerations suggesting that Ben is likely to be correct. There's a clear sense in which, taking all of Alison's evidence at "full strength," Ben's dissent could quickly and simply be dismissed, no matter how much evidence Alison had about Ben's greater reliability in just this sort of problem.[4]

For this reason, various writers have argued that what's rational for agents to believe in disagreement situations depends in part on some assessment of the probative force of the disagreement that's independent of the particular reasoning in question.[5] At this point, these "independence principles" have been quite vague, and real difficulties arise in thinking about how to make them precise. For example: how exactly are we to delimit what must be put aside or bracketed in supporting the agent's independent assessment of how likely it is that her friend has made a mistake this time? But I want to put these difficulties aside here. If some sort of independence requirement will be part of the right theory of rationally accommodating higher-order evidence, perhaps we can make progress on other fronts by postponing those worries, and just supposing that some sort of independence requirement applies. So I will concentrate on a different question about rational requirements in disagreement situations: Supposing that our agent forms some suitably independent reliability assessments of herself and her friend, *how in general is this assessment supposed to figure in constraining rational belief*? In the next section, I'll look at one answer to this question about disagreement, and at how it might be extended to answer similar questions about other sorts of higher-order evidence.

2. The Simple Thermometer Model (STM)

In thinking about the Logic Disagreement case, one natural answer to our question would see the relevant rational requirement on Alison's post-disagreement credence in terms of two main bits of evidence: Ben's credence, and her own initial credence. The idea is that both of these

---

[4] This is why even somewhat conciliatory views—e.g., Thomas Kelly's (2010) "Total Evidence View" or Jennifer Lackey's (2010) "Justificationist View"—will have to employ some form of independent evaluation of the probative force of the other's disagreement. See Christensen (2011) for more discussion of this point.

Kelly (2013) resists this argument on the grounds that deductive reasoning does not justify absolute certainty, even absent disagreement. But this is compatible with the plausible view that in ordinary, disagreement-free situations, we typically do have higher-order reason to have less than complete confidence in our reasoning. If taking such considerations into account generally requires independent assessments, then independent reliability assessment would figure into explaining our lack of rational certainty even in disagreement-free situations.

Moreover, rejecting any sort of independence requirement leaves a key fact unexplained. The Quick argument would be an excellent way of dismissing Ben's dissent if, for example, Ben's opinion was just an ordinary inductive indicator of ~C. If Ben did not know P1 and P2, and his low credence in C was just due to his knowledge of statistics to the effect that C was only true for 1/5 of the parties that took place in Alison and Ben's social group, it would seem that the Quick Argument would easily justify Alison in retaining very high confidence in C. There's more to be said here, but it will have to wait.

[5] See, for example, Christensen (2007a, 2011), Elga (2007), Kornblith (2010).

psychological states would be taken as indicators of the truth of C (hence the thermometer metaphor). The way in which these two indicators constrain Alison's final credence would be determined by Alison's dispute-independent assessment of her own and Ben's reliabilities. So, for example, if Alison's independent assessment has herself and Ben being equally reliable, STM might require that her final credence roughly split the difference between her own initial credence and Ben's. [6]

By having the agent treat herself and her friend as thermometers, this sort of model nicely captures the insight that rationally responding to disagreement requires avoiding a certain sort of question-begging reliance on one's own reasoning. The model also runs into intuitive difficulties. But before trying to state the model somewhat more carefully, and before delving into its problems, let's see how the basic idea might be extended to other sorts of higher-order evidence cases. We can begin with an example:

> **Logic on Drugs:** Carmen is told P1 and P2 by an impeccable source. Carmen thinks about these facts, and becomes highly confident that C. Then Carmen learns that before she started to think about the party, someone slipped her a powerful drug. The drug distorts people's truth-functional reasoning about complex social situations. It causes people doing this sort of reasoning to reach incorrect conclusions in 20% of the time. Carmen has played with this drug before, at parties. She has a long history of reaching wrong conclusions 20% of the time, even while feeling as if she's perfectly clear-headed. She reflects on all this, and becomes less confident in C—say, she reduces her confidence to around .8. I will assume that this is the most rational credence for her to adopt.

To explain how this is rational, we have to invoke Carmen's independent reliability assessment of herself. [7] For example, the more powerful the drug tends to be, the less Carmen should trust her own reasoning. So we have to ask a question analogous to the one we asked about Alison in Logic Disagreement: in cases like this, how in general is Carmen's independent assessment of her own reliability supposed to figure in constraining her rational credence?

In our particular example, where Carmen's initial credence in C is high, the intuitively right result is lowered credence, with the degree of lowering corresponding to the strength of the drug. But lowering credence can't be the general prescription. For example, consider an agent forecasting the weather who reaches .6 credence in rain tomorrow. Then she gets evidence that her forecasts are often distorted by wishful thinking. And she's planning a picnic for tomorrow. Here, it seems that this evidence should push her credence higher, not lower.

---

[6] The basic idea here, and the way STM is developed below, is very similar to Elga's (2007) Equal-Weight View of disagreement. See White (2009) for an extended development and critical discussion of the thermometer idea. Also see Enoch (2010) for related discussion.

[7] Why "independent"? If it were rational for Carmen to base her reliability assessment on her reasoning from P1 and P2 to C, an analogue of the Quick Argument above would allow her to dismiss the possibility that the drug had distorted her reasoning.

One attractive idea here is to extend the thermometer approach to Carmen's sort of case. Carmen would simply treat herself like a thermometer—that is, she would treat her initial credence as an indicator of the truth of C. The import of this indicator would be given by Carmen's independent assessment of her reliability. On this view, Carmen's final credence in C should match the credence in C that would be rational given Carmen's initial credence in C, interpreted as an indicator of C, via Carmen's independent reliability assessment of herself. It's plausible that something along these lines would yield the reasonable verdict that Carmen should have about .8 confidence in C.

Now I realize that these descriptions of STM have been extremely vague. So I want to try to give a slightly less sketchy sketch. In particular, I want to be a little bit clearer on what a reliability assessment might look like, and how it might be applied to agent's initial credences to yield a final credence.

First, it seems that the relevant sort of reliability assessment cannot be represented as a simple number that captures a person's general tendency to get things right. It should be useable when the agent gets evidence that she tends to be overconfident, or underconfident, or unreliable in other, more complicated ways. A thermometer can be useful if one has reason to believe it reads too high when in direct sunlight, or too low in the range below 20 degrees. Similarly, a person's credences can be evidentially useful if one has reason to believe that she's often underconfident in unpleasant prospects, or overconfident when hungry. And so on.[8]

A second point is about what sort of notion reliability is. Some people who have discussed this sort of model have understood reliability in terms of objective chances, or expectations of objective chances.[9] And perhaps it is right to think that there is an objective chance of my reacting in certain particular ways when I'm hungry, and other ways when I'm hungry and mildly tired on a sunny day. I'm not sure. But it would be nicer not be committed to the claim that these objective chances exist. And I don't think that conciliatory responses to higher-order evidence should require such objective chances to do any significant work. (For example, I think that someone who believes that nontrivial objective chances require indeterminism, and who also believes that determinism is true, should still take higher-order evidence seriously.) So I'd prefer to leave objective chances out of our understanding of the sort of reliability assessments that figure into STM.

Instead, we might understand the independent reliability assessments in terms of hypothetical credences. For example, in Logic on Drugs, the following might be true of rational Alison: bracketing her specific reasoning from P1 and P2 to C, her hypothetical credence in C, given that her actual initial credence in C was very high, that she's a very reliable thinker, that she's been drugged in a certain way, etc., is .8. As noted, this way of thinking about reliability is very close to that employed in Elga's (2007) account of disagreement evidence. But while Elga put his proposal in terms of conditional probabilities, I want to avoid taking these hypothetical credences to be literally

---

[8] For these reasons, knowing a person's level of credence in P is more informative than just knowing which of P or ~P the person has more credence in. I believe that the way STM takes the former into account gives it an advantage over Sliwa and Horowitz's "Guess Calibration" or Schoenfield's "J-Calibration," which work with the latter. I'm not fully sure of this, though, as some of the related issues are complicated.

[9] See White (2009), Sliwa and Horowitz (2015), and Schoenfield (2014).

probabilities, because probabilities bring in problems of logical omniscience. However, I think we can understand these credences in a way that doesn't assume that they're probabilistically kosher.

If we can think of the relevant sort of reliability-assessments in terms of these hypothetical credences, then STM will look something like this, when applied to Carmen's case:

> Carmen's final credence in C should match her *independent hypothetical credence* in C: that is, her credence in C, independent of her reasoning from P1 and P2 to C, but taking into account the fact that she reached her initial very high credence in C.

It's understood that while this hypothetical credence will bracket Carmen's reasoning from P1 and P2 to C, it will take account of not only her having reached very high initial credence in C, but other evidence that's relevant to her reliability. This would include, of course, the information about the drug. And in other cases, it would include information about fatigue, wishful thinking, implicit bias, etc.

In general then, we arrive at the following (somewhat-less-vague) model for accommodating the sort of higher-order evidence Carmen gets in the Drugs case:

> **Simple Thermometer Model (STM)**: in cases where the agent A has reached an initial credence in C, and then gets some higher-order evidence, A's final credence in C should match her independent hypothetical credence in C.

It seems to me that that this general model of accommodating higher-order evidence should be able to cover disagreement cases as well. The disagreement of others can be seen as basically providing specific information relevant to the reliability of my thinking on the disputed matter. In typical cases, when a friend is more confident in C, it's evidence that I am underconfident; and vice-versa. And the strength of this evidence varies with my evidence about the reliability of my friend. So the idea would be that the disagreement would bear on an agent's final credence by way of figuring into the reliability-assessment the agent should make about her own thinking on the relevant topic (just as the disagreement of a friend's thermometer may bear on the reliability assessment I should make of my own thermometer).[10]

This would make sense of the sorts of "weighted average"-style conciliatory verdicts that often seem right in cases discussed in the disagreement literature. For example, consider an ordinary case where I rationally reach credence .3 in C, and my equally-informed friend has credence .7. If I have excellent independent reason to think she's equally reliable on this issue, the weighted-average verdict is that I should (at least roughly) split the difference. But if I have reason to think she's generally much more reliable than I, the weighted-average verdict is that I should move my credence closer to .7. I think that STM should be able to give similar results, since in the second case the disagreement serves as stronger evidence that my credence is too low.

---

[10] Seeing disagreement in this way goes most naturally with one of the two main ways "peerhood" has been understood in the disagreement literature. Some of the literature sees peerhood in terms of rationality, and some sees it in terms of reliability. These parts of the literature are not wholly distinct, as some of it (for example, Christensen (2007a)) confuses the two notions. The present approach goes most naturally with the reliability notion. See Schoenfield (2014) and Christensen (2014) for more on the implications of this issue.

I also think the STM could make sense of cases where the simple weighted-average idea does not seem to give the right results. So consider a case where I disagree with five other reliable agents. It may well matter whether these agents tend to think alike—so their mistakes would tend to be correlated. Suppose there were several ways of thinking about the issue in question, and I had information about whether the five are likely to have thought about the issue in similar, or dissimilar, ways. If the other five were very similar reliable thinkers, it would seem that I had weaker evidence of my own unreliability. And this could obviously affect the amount of revision rationally required.[11]

Finally, there are cases where no sort of averaging seems to be correct. For example, suppose that Alison in the Logic Disagreement story rationally forms .98 initial credence in C, then finds out that Ben formed credence .97. Here, it seems that learning of Ben's credence should cause Alison to be more confident that C is true, not less, as an averaging account would have it. For these reasons, it seems to me that if something along the lines of the STM could work, it might help us unify our treatment of disagreement cases and other cases involving higher-order evidence.

There are, however, a number of (at least apparent) problems with STM. Let us turn to examine three of them.

3. Worries about STM

a. First Worry: Satisfying STM requires having irrational credences

One version of this worry has been brought out very clearly by Kelly in the disagreement debate.[12] The problem cases involve agents who form *irrational* initial credences, then follow the model. For example, think about Ben in Logic Disagreement, who formed an irrationally low initial credence in C. If he compromises with Alison's credence, satisfying STM, his middling credence still doesn't seem rational. After all, P1 and P2 entail C. And Alison's high initial credence, if anything, gives Ben more reason to be highly confident in C. So satisfying STM seems to require having an irrational credence in this case.

And as Kelly shows, this problem is especially vivid in cases where both agents initially screw up in the same direction. So consider a variant case in which Ben and Bob reach different irrationally low initial credences in C. STM would have them end up with a compromise low credence. But it's pretty implausible that that this compromise low credence would be rational.

It's also worth noting that the same problem occurs in one-person cases. Consider Dan, who is in a situation just like Carmen's in Logic on Drugs. But Dan botches his initial reasoning completely. He forms an irrational extremely high initial credence in ~C. He then takes account of the drug information, and treats himself like a thermometer, scaling back his confidence in ~C a bit—say, to .8. Here, it does not seem that Dan's high credence in ~C is rational.

---

[11] There are interesting complexities here; see Lackey (2013), Barnett (ms).

[12] See Kelly (2010). Others have discussed the problem as well, including Schoenfield (2014) and Sliwa and Horowitz (2015). Cohen (2013) sketches a way in which this sort of model might be defended against the objection, at least in cases where the agent's mistake is not egregious.

So, how bad is that problem? First, I'm inclined to accept Kelly's insight here: the final beliefs of these agents are irrational. So these cases show that satisfying STM is not sufficient for rational belief.

But I'd also point out that rational requirements need not be shown wrong by this sort of result. This is a familiar point made in thinking about requirements on agents who have in some way failed rationally.[13] So consider Jocko, who irrationally believes that all black cats bring bad luck. He meets Fluffy, and comes to believe that Fluffy is a black cat. Now it may be a rational coherence requirement that:

If you believe that x is an F, and you believe that that all Fs are Gs, you believe x is a G.[14]

Holding fixed Jocko's irrational belief that black cats bring bad luck, and his belief that Fluffy is a black cat, the only way Jocko can satisfy the coherence requirement is by believing that Fluffy brings bad luck. If he does this, of course, this belief won't be rational. But that doesn't show the coherence requirement incorrect.

With this in mind, it's worth noting that Kelly-style cases involve agents who've already formed irrational initial credences in the relevant propositions. And on the most natural interpretation of these cases, the agents' failures to assess the import of their first-order evidence correctly persist. So the irrationality of their final credences may not flow from their reacting improperly to their higher-order evidence. It may flow from their continuing irrational misreading of their first-order evidence.

We can put this point in terms of our thermometer metaphor: Rationality does not just require properly treating oneself as a thermometer. It requires properly *being* a thermometer as well. STM is an instance of the former requirement, but it does not speak to the latter.

I think that this point is strengthened by thinking of another sort of case where agents form irrational beliefs by reacting to higher-order evidence as STM recommends. This second sort of cases have not received the kind of attention Kelly-style cases have, but they seem illuminating. Consider Dex, who is about to look at the logic problem we've been considering. Suppose that Dex knows of the existence our special reason-distorting drugs. And suppose he also has a long track-record of making mistakes in 20% of cases where the drugs were given to him. But Dex is affected by a strong case of epistemic arrogance. He thinks he's too intellectually strong to be affected by drugs; he puts his own bad past performance under the drug down to repeated bad luck. So, before looking at the problem, Dex's credence in any answer he arrives at, even conditional on his having been drugged, is very high.

---

[13] Similar points are often made by supporters of "wide-scope" rational requirements. The point that satisfying a particular model for rational response to disagreement need not mean that the agent's resulting belief is fully rational is made in Christensen (2011) in response to Kelly (2010). See Cohen (2013) and Schoenfield (2014) for more discussion of this issue.

[14] I actually doubt that this is a rational requirement for categorical belief, for reasons related to preface cases. But there are requirements that are close enough, and the differences will not matter here.

Suppose Dex is given the problem, thinks about P1 and P2, and becomes highly confident in C. He then is told that that he was drugged before thinking the problem through. Now let's also suppose that Dex has come under the influence of conciliationist pro-STM propaganda. So he wants to treat himself like a thermometer. But of course, this entails interpreting his initial credence according to his independent reliability assessment of himself. And, as we've seen, his independent estimate of his own reliability is pretty high. So, despite the evidence about the drugs, Dex remains extremely confident in his initial answer.

To my mind, this is another sort of example where someone follows STM, and still ends up with an irrational belief. Dex's case involved an extreme sort of epistemic arrogance. Perhaps a more common kind of epistemic arrogance occurs when people consider most others their epistemic inferiors, even independent of any disagreement, and in spite of ample evidence that the others are equally likely to get things right. But in general, the issue will arise in cases where agents have irrational self-assessments, for whatever reason.

And again, we don't have to see such cases as constituting a problem with STM; they really help show that satisfying STM should not be expected to be sufficient for rational belief. STM takes the agent's initial credence, and her independent reliability assessment, and uses those to constrain the agent's final credence. But when the inputs to the model are irrational, the output can inherit that irrationality. And this does not show that STM is itself defective—just that it's incomplete.

b. Second Worry: STM leaves certain agents epistemically screwed

Suppose that we allow that STM provides only a necessary condition for rational belief, and thus is not shown wrong by failing to prohibit Dan's or Dex's irrational credences. Still, one might think that there was something more disturbing about the way STM applies to their cases. On the one hand, the credences they arrive at in accordance with STM are not fully rational. On the other hand, one might worry that if STM is a rational requirement, then any other credence they adopted would also violate a rational requirement! So they'd seem to be in a way epistemically screwed.[15]

In assessing this worry, I think it's helpful to think back to Jocko and the black cats. In particular, let's ask whether Dan or Dex is faced with a problem more severe than the problem that faces Jocko. In the black cat case, perhaps Jocko actually does have an option that doesn't violate any rational requirements. He could satisfy both the coherence norm and respect for the evidence by giving up his irrational belief about black cats. He may not do this. But the fact that he can't satisfy all the rational requirements *while continuing to believe that black cats bring bad luck* should not worry us at all.

Now I think it's plausible that Dex's situation is similar to Jocko's. If we think of all the evidence he has, he seems to have a similar option: He should give up his irrationally arrogant self-assessment. We might imagine him becoming suddenly rational, and realizing that the evidence requires a less sanguine assessment of his reliability. If he did that, and then formed a new credence in C in accordance with STM, it would seem that his final credence would be as rational as

[15] See Schoenfield (2014) for a closely-related objection to a relative of STM. For worries about this general sort of issue see Schechter (2013), Lasonen-Aarnio (2014) and Titelbaum (2015).

Carmen's—in fact, the credence he reaches might be just the same as the one she reaches, and it would be based on analogous considerations. So it doesn't seem that Dex's case poses any special problem for STM.

Things are not obviously that simple for Dan. Suppose we ask: given Dan's evidence, what set of beliefs would be most rational for him to have? To make things vivid, we could imagine Dan, too, suddenly becoming rational. So the first thing to notice is that P1 and P2 would no longer seem to Dan to support ~C, as they did in our original story; presumably, he'd give up his initial credence for reasons entirely separate from his higher-order evidence about the drug. We might even imagine him forming a provisional credence in C based on just his first-order considerations, before adjusting for his higher-order evidence.[16] Presumably, this provisional credence would match Carmen's initial credence. Does this mean that he, too, would be in a situation essentially like Carmen's?

To make the strongest case for this possibility, we might imagine that if Dan were fully rational he'd have an "aha!" moment. In this moment, we might imagine him seeing exactly where he went wrong in his initial thinking about P1 and P2. And it seems clear that having this sort of experience would support giving his initial credence less weight. Would this make his initial credence completely insignificant, so he'd be in an epistemic situation parallel to Carmen's? If it did, then maybe we could treat both Dan's and Dex's cases like we treated Jocko's.

Unfortunately, I don't think that all versions of Suddenly Rational Dan's case will work out so neatly. In part, this will depend on how we fill out the details. We should recognize that it's consistent with having an "aha!" moment that one get things wrong after all. So to fill in details in a way most likely to cause trouble for STM, we might imagine Dan remembering a long series of misleading "aha!" moments while drugged. What credence would be rational for Dan in that sort of case?

We can see Dan in this sort of case as in a sense disagreeing with his past self. We also saw that a friend's disagreement can affect the credence that STM recommends. It does this by feeding into the independent reliability assessment the agent should make. This is how STM can explain why Alison should lower her credence in cases where Ben disagrees with her. But this suggests that when we apply the model to Dan's situation, we should treat Dan's initial credence as evidence against the reliability of his provisional credence. If he has in fact had a series of misleading "aha" moments when drugged, then his initial credence will still be quite evidentially significant. If that's right, then it would seem that STM would recommend a final credence in C for Dan that's significantly lower than the one Carmen ended up with. So Dan's situation would not be just like Carmen's, and the easy analogy to Jocko's case disappears.

Does that mean that Suddenly Rational Dan is in a position, vis-à-vis the rationality of his final credence, that's worse than Carmen's? I think that answer, at least arguably, is "no". Dan's situation is different from Carmen's in that Dan has a bit of evidence—his own low initial credence in C—that Carmen lacks. So his ending up with a credence different from hers need not itself mean

---

[16] One might worry that this provisional credence wouldn't really be a credence, or, more directly, that Suddenly Rational Dan may lose his initial credence without forming a provisional credence based on just P1 and P2. Seen this way, the case will be an example of the problem discussed in the next subsection.

that he falls more short rationally. If that's right, then there may well be an option for Dan that is just as rational as the most rational option for Carmen. Insofar as that's right, Dan, like Dex, would pose no special problem for STM.

But I should also say that I can see reasons for resisting this verdict. After all, the bit of evidence that's lowering Dan's final credence is his own initial credence. And that was the direct product of his own rational blunder. So, one of the rational determinants of his final credence is his own irrational initial belief. For this reason, some may be inclined to deny that his final credence is rational. If that's right, then Dan may be in a situation importantly different from Jocko's.

I'm not sure what to say about this. But it is worth pointing out that even if Kelly-style cases do leave Dan epistemically screwed, in the sense of having to violate some requirement of rationality, it's not clear that this poses a particular problem for STM that's not already present in any view which takes higher-order evidence seriously. That's because it's plausible that even agents such as maximally rational Carmen are faced with violating some requirement of rationality. It's plausible that logic supports a rational requirement that one be at least as confident in C as in (P1 & P2). If Carmen is required to lose confidence in C while leaving her confidence in (P1 & P2) intact, then it seems to me that we already have conflicting rational requirements. So I think that conflicting rational requirements come with the territory, when one takes higher-order evidence seriously as a constraint on rational belief.[17] If that is right, then the main thing that Kelly-style cases show about STM is what we saw in the previous section: that it's incomplete, in the sense of not providing a sufficient condition on rational credence.

c. Third Worry: STM is incomplete in a more serious way

The third worry about STM does not involve problems in applying it to cases where agents have adopted irrational initial credences. Instead, it involves agents who don't have the credences that STM takes as inputs. Most obviously, there are cases where agents have not formed any initial credence on the basis of a batch of first-order evidence before they get the higher-order evidence.

For example, consider maximally rational Elena, who goes to solve our logic problem *knowing in advance* that she's been drugged. Since she knows she's been drugged, Elena won't become extremely confident in C, even though it seems to her that P1 and P2 entail it. So she never forms an "initial credence"—that is, a credence that's formed independently of the higher-order evidence about the drug. So STM simply doesn't apply to her case.[18]

A similar problem will obviously apply to an agent who is missing the other input to STM: the independent reliability assessment. So it looks like STM suffers from an incompleteness that's more disappointing than the one revealed by Kelly-style cases. Kelly-style cases show that even if

---

[17] Some may see the predicament that Carmen ends up in as more theoretically problematic than the one Dan faces. At least in the moral case, some have been attracted to the view that the correct account of morality should only allow conflicting moral requirements to arise in situations agents have gotten themselves into through their own prior moral failure. See Christensen (2007b, 2010) for defense of the picture on which epistemic conflicts will arise even for those innocent of prior epistemic sin.

[18] This basic point also applies to Elga's (2007) account of disagreement, in cases where someone learns of her friend's opinion before confronting the first-order evidence.

STM correctly captures the constraints that higher-order evidence places on belief, it does not yield sufficient conditions for rational belief, because it doesn't account for other constraints evidence places on rational belief. But the case of Elena seems to show that STM cannot even capture the constraints that higher-order evidence places on beliefs in some cases. Perhaps this does not show that it's wrong as far as it goes. But since the cases where it does not apply are cases in which we want to know how an agent should take higher-order evidence into account, it would obviously be nice to have an account without this limitation. In the remainder of the paper, I'll examine an account that strives for greater generality.

4. The Idealized Thermometer Model

In thinking about how Elena should react to her higher-order evidence, an attractive way to begin is with the supposition that Elena should end up with the same credence Carmen ends up with. That's where she'd be if she'd gotten the same evidence, but in a different order, and had formed the rational initial credence on the basis of P1 and P2, and then followed STM when she learned about being drugged. In effect, we'd be substituting the credence that's *rational* given P1 and P2 for the credence the agent *actually forms* on P1 and P2. [19]

We can make a similar move with respect to the independent reliability estimate: invoking the assessment it would be *rational* for the agent to make where STM would take account of the agent's actual assessment. Thinking along these lines suggests a requirement along roughly the following lines: Elena's final credence in C should match the credence in C that would be rational, given an agent like Elena's having the credence in C that would be rational given her first-order evidence, when that credence is interpreted as an indicator of C, via the independent reliability assessment it would be rational for Elena to have of herself. If we spell out the general idea in terms of hypothetical credence, as we did for STM, we get something like this:

> **Idealized Thermometer Model (ITM):**
>
> Let's use *n* to stand for the credence in C that would be rational given C's first-order support by the agent's evidence.
>
> Then the credence in C it would be rational for the agent to form, given *all* her evidence, is the credence in C that would be rational:
> - independent of C's support from first-order considerations,
> - conditional on a relevantly similar agent adopting credence *n* in C on the basis of first-order support from the agent's evidence.

[19] This move is analogous to Sliwa and Horowitz's move from Guess Calibration to their Evidential Calibration. One difference is this: Evidential Calibration applies a reliability assessment to whichever of P or ~P the agent's first-order evidence rationally supports more strongly (if either). ITM applies such an assessment to the level of credence in P that would be rational given first-order support from the agent's evidence. ITM is thus directly sensitive to degree of first-order support in a way that Evidential Calibration is not. See Sliwa and Horowitz (2015), §5.2 for discussion of this issue.

> By "relevantly similar," I mean someone who is similar with respect to the reliability evidence the agent has about herself. So the evidence that this hypothetical credence is based on would include evidence about the agent's track record, exposure to drugs, the presence of disagreeing agents, etc.

Something like ITM seems a natural way of extending the insights behind STM to cases where agents don't have the initial credences that STM takes as inputs.

Moreover, I think that ITM also addresses a different limitation of STM. In thinking about the worry that Kelly-style cases seemed to pose for STM, we saw that STM required agents to treat themselves as thermometers, but that this was not sufficient for rational belief. One way of putting it was to say that rationality requires not only treating oneself as a thermometer, but also requires properly being a thermometer, where properly being a thermometer involves reacting rationally to one's first-order evidence. Now ITM puts *rational* credences where STM had the agent's *actual* initial credences. In doing this, it effectively incorporates (at least part of) the requirement that agents be good thermometers.

Thinking about ITM this way suggests that it's not just a companion to STM, useful because it applies to cases where STM is silent. It may offer a more complete assessment of the rational requirements on even agents who have formed the relevant initial credences. So, we might notice that ITM, unlike STM, does condemn the credence that Dan ends up with in the Kelly-style case, when he adopts .2 credence in C. ITM seeks, in a way that STM does not, to capture the way that first-order evidence and higher-order evidence interact in constraining rational credence.[20] And it does so in a way that avoids a worry mentioned above about views like Kelly's Total Evidence view: since the first-order and higher-order evidence don't simply get "weighed" against each other, ITM makes clear how higher-order evidence can play a role in constraining one's credence in C even in cases where one's first-order evidence entails C. ITM would also condemn the credence arrogant Dex ends up when he remains extremely confident in C despite the drug evidence, due to his irrationally high assessment of his own reliability.

For these reasons, ITM seems to me to be worthy of exploration as a more general account of the rational requirements imposed by higher-order evidence.

Nevertheless, there are a number of worries one might have about ITM—some, no doubt, that I haven't seen. In the remainder of the paper, I'll look at four of the most interesting worries that have occurred to me.

5. Worries about ITM

a. First Worry: ITM ignores important evidence.

The first worry centers on a factor that is crucial in constraining rational credence in STM, but which seems to have dropped out of sight in ITM: the agent's initial credence (in cases where these credences exist). To make the problem vivid, I'll use a kind of example due to Dominik Berger (ms). Consider a variant of our Logic Disagreement case where Ben begins with very good dispute-

[20] Sliwa and Horowitz make a parallel point about their Evidential Calibration.

independent reason to believe that he's a *much better* logician than Alison. In particular, he has a long track record indicating that when he and Alison disagree in thinking about logic problems, he's almost invariably the one who's right. Today, however, he blunders and forms a very low initial credence in C, despite being extremely confident of P1 and P2. He then learns that Alison is quite confident in C on the basis of P1 and P2.

One might worry that if ITM has replaced the agent's actual initial credence with the initial credence it would be rational for the agent to have, then Ben's having reached his low initial credence would become irrelevant. Given that the rational initial credence in C is very high, and Alison's high initial credence in C would only seem to support the reliability of that high credence, one might think that ITM would say that Ben should be extremely confident in C. And this might seem wrong: after all, Ben has excellent reason to think himself a highly reliable reasoner in the relevant sort of matter. If a *different* highly reliable reasoner had reached low confidence in C on the basis of P1 and P2, ITM would certainly require Ben to take that into account. That's because disagreement evidence is like drug evidence—it's figured into the independent reliability assessment, as it is in STM. So Ben's final credence in C should in general take into account the fact that a highly-reliable logician became highly confident in ~C on the basis of P1 and P2. The fact that this highly reliable logician happens to be himself should not disqualify it as useful evidence. [21]

I think it's clearly right that Ben's low initial credence should have a significant effect on the credence it's rational for him to end up with. But I think also think that ITM, properly understood, will give this result. A relevantly similar agent in Ben's epistemic position, who now forms the rational (high) first-order-based credence in C will have, as part of his reliability-relevant evidence, the information that he initially had formed a low credence in C. And this, just like the information that a different agent had formed a low credence in C, will affect the reliability assessment that ITM applies to the rational first-order-based credence in C. In particular, it will have the effect of moving the final credence Ben should adopt closer to his initial credence than it would be if he had never formed that initial credence. So Ben's initial low initial credence does not drop out of the picture after all.

b. Second Worry: ITM is not followable

The second worry I want to examine is similar to one that has been brought up as a criticism of "Right Reasons" views of disagreement (e.g. Kelly (2005), or even Kelly's more moderate (2010) "Total Evidence" view. On these views of disagreement, the requirements on agents in disagreement situations may be asymmetrical, even when the agents have very strong dispute-independent evidence of equal reliability. In a case like Logic Disagreement, where Alison has initially evaluated the first-order evidence correctly, and Ben has evaluated it incorrectly, these views say that Alison is required to compromise less, or not at all, with Ben. But Ben should move significantly closer to Alison's initial credence. People sometimes object to these views that they're not followable in a

[21] One of Berger's targets is Christensen (2011, 6-7) which suggests that the important determinants of rational belief in disagreement situations are the first-order evidence and the other agent's credence, and that "the first-person psychological evidence is relatively inert". I now think that that can't be quite right. See also Schoenfield (2014) for a related, but different, way of raising difficulties in this neighborhood.

certain sense. After all, when one is actually in such a disagreement situation, it seems that one can't tell whether one is in Alison's position, or in Ben's. So one can't use the theory to decide whether, or how much, to compromise.[22]

One might make a similar complaint against ITM. ITM applies a reliability assessment to an agent's having the credence that the first-order evidence actually supports. But if one is unsure what that credence is, the objection would go, one can't use the theory to determine what credence it is rational to end up with. As Baron Reed pointed out to me, it might even be particularly hard for an agent to assess what the first-order evidence supports when she has higher-order evidence as well. Since, as psychologists tell us, expectations affect cognition, an agent's assessment of the first-order evidence's import may be distorted by her concerns about her own thinking on the relevant matter.

Now I think that Kelly has made an important point that's relevant to this general sort of criticism. The same sort of phenomenon exists, even if we put aside higher-order evidence. As Kelly (2005, 180) says, "[H]ere as elsewhere, life is difficult. On any plausible account of evidence, we will be extremely fallible with respect to questions about what our evidence supports." But it's presumably nonetheless a requirement of rationality to have the credence supported by one's evidence. So acknowledging that we can't always get evidential relations right, or be certain that we have gotten them right, is compatible with holding that failure to get evidential relations right is a rational failure. In epistemology, as in ethics, the correct norms may put perfection beyond what ordinary humans can generally achieve. So I think that we should not reject ITM simply on the grounds that it's not followable in some very strong sense.[23]

c. Third Worry: ITM denies Elena doxastically rational credences

A related worry centers on the doxastic notion of rationality. Presumably, if an agent arrives at the ITM-sanctioned credence by guessing, her credence will still not be *doxastically* rational. In order for Elena's credence to be doxastically rational, it's plausible that it must be produced or sustained by the factors that make that credence propositionally rational. So although Elena never formed a credence on the basis of P1 and P2 alone, her final credence would have to be produced or sustained in a way that was sensitive to the degree of first-order support that P1 and P2 actually give to C. If that seems impossible, then one might worry that ITM-sanctioned credences could not come out doxastically rational, even for maximally rational agents in Elena's kind of situation.

I don't think that it really is impossible for Elena's credence to be causally sensitive in the appropriate way to the factors which make it rational on ITM. Of course, the causal-inferential relations responsible for doxastic rationality need not—and typically will not—be conscious. But we

---

[22] See, for example, Enoch (2010). Schoenfield (2014) takes this sort of point to undermine a possible line of motivation for the relative of ITM that she discusses.

[23] But might some degree of followability be part of any adequate account of rational belief? Kelly's point may help us resist very strong followability requirements, but one might still worry that ITM was not followable *enough*. See Leydon-Hardy (this issue) for a nice presentation of this sort of worry. At this point, it's not clear to me exactly what sort of followability requirements should constrain our theory of rational belief. So I'll leave this issue unsettled, with the hope that ITM is followable enough to be worth exploring further.

might be able to throw some light on the problem by thinking about how maximally rational Elena might consciously go about trying to form an accurate credence in C.

First, she would notice (correctly, of course) that P1 and P2 seemed to support C. And then she'd use her reliability assessment of herself—informed by her information about the drug—to treat her first judgment as an indicator. In our case, she would reduce the high credence that P1 and P2 seemed to her to support, to a lower credence in C. And being maximally rational, she'd presumably get that part right, too.[24]

These things don't seem to be impossible to do consciously. And if that's correct, then a well-functioning agent may after all be sensitive to the right things—the factors responsible for a certain level of credence being maximally rational in her situation. So I see no clear obstacle to Elena's forming a doxastically rational credence in C, even if ITM is true.

d. Fourth Worry: ITM breeds akrasia

The final worry I want to discuss here concerns an agent's attitudes toward her own ITM-sanctioned credences. As we saw, it's not impossible for Elena, in trying to believe accurately, to succeed in forming the credence ITM prescribes. But it's worth thinking about how she might regard herself when she has done this—especially if we keep in mind the insight behind the above worry about followability. The insight there was that even those who comply with ITM precisely cannot be sure that they are doing so. And this of course applies to Elena, even if she is thinking exactly as ITM requires. At the end of the process, it seems clear that she should have significant doubts about whether she has complied with ITM. After all, as she can plainly see, she has in fact complied with ITM only if her assessment of the first-order evidence was rational—that is, if it was not distorted by the drug. But the whole rationale for her reducing confidence in C is that she should give significant credence to the possibility that her appreciation of the first-order evidence *was* distorted by the drug. And if it was, her current credence in C is too high.[25]

Now this might not seem like much of a worry. But the potential worry can be sharpened using a kind of example Schoenfield (2014) has deployed in criticizing Sliwa and Horowitz's

---

[24] There is nothing too strange, I think, about the idea that an agent can be rationally bound by a condition that makes reference to what some portion of her evidence supports, without going so far as to form a credence on the basis of just that evidence: Consider a detective who discovers that the mob's henchperson has gone around planting misleading evidence—say, evidence that would (considered on its own) strongly suggest the guilt of someone who is in fact innocent. And suppose that the evidence she discovers (considered on its own) strongly suggests that Lee is guilty. It would seem that our detective should, and could, form her credence as follows: She would not form an initial credence in conformance with what her ordinary forensic evidence supported—she would never believe Lee guilty. But her final assessment would have to be sensitive to what that evidence supported, so she would end up rationally confident that Lee was innocent.

[25] Sliwa and Horowitz (2015, §4.4) use a similar example in acknowledging that their Evidential Calibration principle would require violations of Rational Reflection, which is a kind of anti-akratic principle.

Evidential Calibration principle. Schoenfield suggests that we consider cases where agents get very good evidence of their *anti-reliability* in forming correct assessments of the first-order evidence.[26]

So let us consider a variant of our Elena story. Suppose that maximally rational Elena gets evidence that she's been dosed with an even more powerful, Extra-Strength version of our favorite drug. People under its influence almost always reach credences diametrically opposite to the ones supported by their first-order evidence when doing complex truth-functional reasoning about social situations. When Elena is then given P1 and P2, ITM says that she's rationally required to form a very *low* credence in C. (The rational credence in C on just P1 and P2 is very high, so treating this high credence as an indicator, when one has excellent reason to think that one has been dosed by a strong anti-reliability drug, yields the opposite, very low, credence.) So let's suppose that Elena forms the requisite very low credence in C. And let's ask some questions about what she should think of her low credence.

First let's ask: should she think her *first-order* evidence supports C, or ~C? I think it's clear that she should think that it supports ~C. In most cases, it's plausible that one should expect that whatever credence is actually supported by one's first-order evidence is more accurate than a credence that's opposite to the one supported by one's first-order evidence. And this is true independent of whether one is in a good position to tell which credence is in fact the one that one's first-order evidence supports. This is particularly clear in the case of logic problems with true premises. So I think it would make no sense for Elena both to have low credence in C and to think that her first-order evidence supported high credence in C. Given that Elena rationally has low credence in C, she should also be confident that her first-order evidence supports low credence in C.

This also seems right from a common-sense psychological point of view. When Elena thinks about P1 and P2, it will seem to her that they entail C. But given her excellent evidence about being drugged, she will not trust that seeming—in fact, she'll take it as evidence that P1 and P2 entail ~C.

So far, this may seem perfectly reasonable. But now let's ask a second question: Should Elena think that her current low credence in C is supported by her *total evidence*?

The first thing to notice is that if Elena thinks that her first-order evidence supports ~C, then she cannot think that her current low credence in C is sanctioned by ITM! That's because, if the first-order evidence really did support ~ C, ITM would say (given Elena's anti-reliability evidence) that she should have *high* credence in C. And it's not just that she can't be *sure* that her credence is the one supported by ITM, as we saw in Carmen's case. Elena must actually be *highly confident* that her credence is quite *opposite* to the one ITM prescribes!

Now suppose that ITM is true and that Elena accepts it. Insofar as she's confident that ITM is correct, she must be highly confident that her own low credence is far below the one supported by

---

[26] I should note that I'm not going to take this sort of example in the same way Schoenfield (2014) does. Part of this may be due to differences between ITM and the model she's criticizing. And part may be due to my being inclined to go a different way with this sort of example. In her subsequent (forthcoming), Schoenfield revisits this sort of example briefly. There, she does not treat it so much as an objection to the ITM-like model she's describing, but as a way of showing that this sort of model is inconsistent with a principle about rational belief which she takes as plausible, but which she thinks may have to be given up on certain attractive views about rationality.

her total evidence. In other words, she is rationally required both to have a very low credence in C, and to be highly confident that her low credence in C is irrational!

So it seems that ITM will require maximally rational agents—at least those who accept ITM—to be very sharply akratic in certain situations. Should we see this as a mark against ITM—or even a fatal defect? It seems to me that, on closer inspection, we should not. In fact, once we see clearly the source of the akrasia ITM requires, we should consign it to the 'feature, not bug' category.[27]

e. In Praise of Akrasia

There is, of course, something that feels a bit unseemly about this sort of epistemic akrasia, at least at first. But it's worth noting that this sort of akrasia will arise on any view that respects a certain sort of verdict about cases where agents get very strong evidence of their own anti-reliability in assessing the first-order evidence. The verdict is this: When an agent has this sort of evidence about herself, what she's most rational to believe in the end is the opposite of what her first-order evidence actually supports. That's because, if she's maximally rational, she'll actually manage to assess the first-order evidence correctly, so that when she takes her anti-reliability evidence into account, she'll end up reversing the correct first-order assessment. So insofar as we find this verdict plausible, we'll hold that if an agent with strong anti-reliability evidence does manage—against what she should expect, given her higher-order evidence—to react to *all* of her evidence as she should, she will up with beliefs that are opposite to the beliefs that are actually supported by her first-order evidence.

Naturally, such beliefs will tend to be inaccurate. And a rational agent who accepts principles of rationality that respect the verdict will be able to see this. So if she has very strong anti-reliability evidence about herself, she'll see that she is faced with two possibilities: either (a) believing something likely to be inaccurate, or (b) believing something irrational. What should such an agent do? She won't aim for (a), since having high confidence that P is too close to having high confidence that a belief that P is accurate. So of course she'll aim for (b): she'll aim for accuracy over rationality.[28]

Now, if she somehow manages, against all odds, to assess the first-order evidence correctly, then she will usually end up with inaccurate beliefs when she also takes proper account of her higher-order evidence. However, she'll also very reasonably have very low credence that she has

---

[27] I should note I'm working with an intuitive notion of akrasia for credences here, according to which being highly confident that your credence in C is way below the credence it would be rational for you to have in C counts as acute akrasia. This is much more intuitively problematic than, say, being confident that your credence in C is different from the ideally rational one, but having no idea whether it's higher or lower. A rough gloss on the notion might be that credence-akrasia involves there being some other credence such that one expects it to be more rational than the credence one has. There are, of course, many different ways to make the rough notion precise. But I take it that the intuitive worry here is clear enough.

[28] Schoenfield (forthcoming) makes a related point in arguing that on accounts of rationality that take higher-order evidence into account, agents should prefer credences rationalized by certain proper subsets of their evidence over credences rationalized by their total evidence.

ended up in that situation, since she'll think it's highly unlikely that she has directly assessed her first-order evidence in the rational way. She'll end up seeing her own belief as likely to be accurate but irrational.

We can see a similar point from a different angle by leaving aside ITM, and thinking back to STM. Let us consider a variation of Carmen's story. As before, suppose that Carmen first learns P1 and P2, and rationally forms an extremely high credence in C. Then suppose she learns that she was slipped the Extra Strength version of our drug. In that case, she should think (falsely, of course) that her initial high credence in C is irrational, and thus likely to be inaccurate. And so she'll adopt a very low confidence in C (as STM would advise), which she'll see as more likely to be accurate.

Now, suppose we take on board Kelly's insight: that is, that a person who begins by having an irrational initial credence, and then corrects according to her higher-order evidence, still has an irrational credence. And suppose that Carmen, being her maximally rational self, shares Kelly's insight. If Carmen reflects on her own situation, she should be confident that her own final credence is irrational, even if she has successfully followed STM. Nevertheless, she'll also be confident that it's accurate, and that a fully rational credence would not have been accurate.

When we see matters this way, the akrasia turns out to be not so implausible after all. The reason, I think, relates closely to an observation made in Horowitz (2013), a paper mostly devoted to arguing that epistemic akrasia is typically irrational. Horowitz discusses a strange sort of case devised by Timothy Williamson, where uncertainty about one's evidence leads to strong akrasia.[29] She points out two differences between Williamson's sort of case and the kind of self-doubt cases she mostly concentrates on. One is that Williamson's cases, unlike usual self-doubt cases, involve uncertainty about what one's evidence is. That won't help us here, because if something like ITM is right, akrasia occurs even in cases where one knows exactly what one's evidence is. But the other difference, which is the one she concentrates on, is more helpful: in Williamson-style cases, one can see in advance that one's evidence will be misleading.

It is obvious in general that rational agents confronted with misleading evidence tend to reach inaccurate beliefs. That's just what we mean by calling certain evidence "misleading". So it's a feature of misleading evidence in general that agents who have such evidence can form accurate beliefs, or rational beliefs, but not both. But in most cases, misleading evidence will lead rational agents to inaccuracy, but not to akrasia, since the agents will have no idea that their evidence is misleading. The strange thing about Williamson-style cases—and, if something like ITM is correct, certain cases of self-doubt—is that they're cases where one's evidence is of a type that's *systematically misleading*.[30]

To see the contrast, consider an ordinary case of misleading evidence: Suppose Theo flips a strange coin ten times, and gets ten heads in a row. He becomes confident it's biased toward heads; but he's wrong: it's a fair coin that happened to land heads ten times in a row. Now Theo just got unlucky—most of the time when agents form rational credences in response to "ten heads in a row"

[29] See Williamson (2011, 2014).

[30] Horowitz, in a subsequent paper (ms), also argues that higher-order evidence should also often be expected to be misleading.

evidence, the credences will be accurate. So these evidential situations are typically non-misleading, since rationality and accuracy are strongly correlated. And as Horowitz points out, this is related to the fact that akrasia in such situations is irrational: if my belief is irrational, it's likely to be inaccurate.[31]

By contrast, anti-reliability higher-order evidence situations are misleading in a non-accidental way: rationality and accuracy are anti-correlated. So being akratic in such situations can be rational, precisely because it does not suggest that one's beliefs are inaccurate. In fact, given that one must expect from the outset that one's belief will be accurate only if it's irrational, akrasia is rationally required.

The lesson about akrasia that we can take from the extreme anti-reliability cases should apply more generally. The fundamental lesson, I think, is this: Epistemic akrasia is not, per se, a problem at all. Thinking that a belief of yours is irrational in a particular way should disturb you—that is, give you a reason to change that belief—only insofar as the particular irrationality indicates that a different belief would have greater expected accuracy.

6. Conclusion/Preview of Coming Attractions

It's worth noting that the sort of extreme anti-reliability evidence we've been imagining is very rare. Most cases of self-doubt—the ones that help make the problem of higher-order evidence seem more pressing—involve much milder evidence of unreliability. So it's worth asking whether the sort of akrasia we see in the extreme anti-reliability cases is peculiar to a few recherché thought experiments.

It seems to me that it is not. The sort of acute self-doubt we see in anti-reliability cases lies at one end of a spectrum—one that will include many more ordinary cases of self-doubt. And just as ITM mandates acute akrasia in anti-reliability cases, it will mandate a spectrum of less severely akratic responses to more ordinary cases.[32]

At the milder end of the spectrum, consider a medical resident who calculates the recommended dosage of a drug, knowing that she's been awake for 27 hours. Or a conscientious manager evaluating resumes, having good reason to think that he's somewhat implicitly biased against female candidates. Or a mother, trying to evaluate whether her child should get the lead role in the school play, while aware that parents tend to overvalue their children's talents. In any of these cases, the first-order evidence might strike the agent as strongly supporting some claim P, but the agent might reasonably reduce confidence in P in response to her higher-order evidence—say from .98 to .8. If the agent accepts ITM, she'll see that her .8 credence is rational if her direct take on the

---

[31] A related point is also made by Elga in his (2013), a paper responding to Williamson cases, and proposing a formal principle describing how a rational agent's credences in general relate to her credences about what credences would be rational for her. Elga puts his point by noting that in Williamson-style cases, the rational credence for some proposition A, on the condition that the ideally rational credence for A is n, will be very different from n. I'll discuss Elga's principle further below.

[32] Thanks to Andrew Rotondo for prompting me to develop this line of thought.

first-order evidence was correct. But she may also be able to see that if it wasn't correct, then ITM requires a lower credence. Since she rationally gives significant credence to that possibility, she has to give significant credence to the possibility that her .8 credence is irrationally high. So very ordinary situations of self-doubt may engender at least mild degrees of akrasia.[33]

Stronger evidence of unreliability can induce correspondingly stronger akrasia. This might occur when an agent is evaluating a complicated social policy that will greatly benefit him, or thinking through a subtle new philosophical argument after enjoying a couple of generous Martinis. And I think that moderately strong akrasia should occur in certain cases of apparent peer disagreement.

For a simplified illustration, think back to Alison, in our Logic Disagreement case. Suppose she arrives at about .5 final credence in C after taking account of Ben's disagreement, and that this is what ITM requires. Alison can see that, given ITM, if her initial take on P1 and P2 was correct, and they support C, then her current middling credence in C is rational. But what if Ben was the one who was correct this time? That is, what if P1 and P2 actually support ~C? In that case, the rational credence for Alison to have in C would be quite a bit lower than .5. (This is because the rational credence for Alison to have in C before taking Ben's disagreement into account would be very low, and the higher-order evidence of Ben's very low initial credence in C would support the reliability of very low credences in C.) Now given our story, Alison quite reasonably thinks it's as likely as not that it was indeed Ben who was right this time. So she should think it as likely as not that her final credence is quite a bit too high.[34]

So given ITM, mild-to-moderate akrasia will be commonly required in everyday life (at least for those who have the correct theory of rational belief). Does this give us more reason to be suspicious of ITM—or even to reject it? I suspect not, though more detailed work on this question will be required. Nothing argued for above tells against the view that there are kinds of akrasia that do indicate significant failures of rationality. In fact, I am persuaded that many instances of epistemic akrasia are quite irrational.[35] The question of whether the cases of akrasia required by ITM are

---

[33] The akrasia in some of these cases may be mild in one dimension, but more severe in another. The medical resident may have relatively low credence that her original judgment was distorted, and thus that her present judgment falls short rationally. But she may also may think that if her initial judgment was distorted, her current credence is quite far from the one rationally required.

[34] While moderate akrasia should arise in many two-person disagreements, it's worth noting that conciliatory responses to large group disagreements among rough peers should be much less sharply akratic. That's because the evidential import of someone in the agent's position forming one credence rather than another will be relatively small, given all the other opinions of parties to the disagreement. So the difference between the credence that's rational if the agent's take on the first-order evidence was correct, and the credence that's rational if the agent's take on the first-order evidence was incorrect, will be small.

[35] This includes the sorts of cases of akrasia defended in Coates (2012), Weatherson (ms.) and Lasonen-Aarnio (2014). See Horowitz (2013) for discussion of this point. It also includes the central case of irrational akrasia discussed in Greco (2014), which uses an expressivist analysis of beliefs about justification or rationality to defend the view that epistemic akrasia is always irrational. Greco notes that those who think epistemic akrasia can sometimes be rational are faced with the task of explaining the appeal of the view that it's always irrational. If the irrationality of typical cases of akrasia traces to their indicating inaccuracy, we

problematic will depend on whether they indicate inaccuracy in a problematic way. Settling this issue is perhaps easier in extreme anti-reliability cases, since they involve such sharp divergences between rationality and accuracy. Settling the issue in milder cases will involve careful attention to what exactly the connections are between rationality and accuracy—or, more generally, what the connections are between a rational agent's credences about which credences are rational in her situation, and her ordinary credences. Elga (2013) proposes a "level-connecting" principle, New Rational Reflection (NRR), which is aimed at providing an account of this relationship, and which is specifically designed to get the Williamson cases right. It is too early to say whether NRR will prove satisfactory in the end.[36] But it's perhaps interesting that a bit of preliminary experimentation with NRR suggests that ITM-induced akrasia is consistent with something like NRR's way of connecting levels. (See the Appendix for a couple of examples.) At this point, however, I cannot say anything more definite.

Nevertheless, I think we can see reason to expect that the correct theory of accommodating higher-order evidence will require at least mild epistemic akrasia fairly frequently. One of the main reasons that higher-order evidence has the epistemic importance it does is that we humans are very fallible in thinking about our ordinary evidence, and prone to making epistemic mistakes pretty frequently. Knowing this about ourselves is useful, but not because this knowledge somehow confers upon us the epistemic grace to see the first-order evidence correctly. Instead, in certain sorts of cases, this knowledge allows us, in effect, to hedge our doxastic bets. But as Kelly has shown us, an agent who gets things wrong and then hedges his bets in response to higher-order evidence does not thereby escape the rational taint of his original error. Given that we all must fairly frequently hedge our bets by taking seriously evidence of our possible unreliability, we must fairly frequently see ourselves as likely to have ended up with beliefs that are tainted. Given the nature of much of this evidence of possible unreliability, we will also often have a good idea which direction we've erred in, provided that we erred. We must, then, often end up with credences that we should see as likely to be irrationally high (or, in different cases, irrationally low). The fact that rationally required akrasia will be fairly common, then, is entirely to be expected. It merely reflects the commonness of situations in which we're rationally required to acknowledge, and to take seriously, ordinary indicators of our own epistemic fallibility.

Appendix

Here are a couple of worked examples showing how the sort of akrasia induced by ITM might interact with Elga's NRR principle. NRR is a coherence requirement which constrains an agent's ordinary credences by relating them to her credences about what credence functions would

---

might well expect that akrasia would strike us as generally irrational. Since rationality and accuracy are generally correlated, reason to think that a belief of mine is irrational is typically reason to think that the belief is inaccurate.

[36] For some doubts, see Lasonen-Aarnio (2015).

be ideally rational in her situation. If we let cr stand for the agent's credence function, and let pr stand for other credence functions that might be ideal, the principle is as follows:

NRR: $cr(A \mid \text{pr is ideal}) = \text{pr}(A \mid \text{pr is ideal})$

The examples below are idealized and vastly simpler than real cases, but this allows us to use numbers tractably. More importantly, my treatment of the examples will cheat in a potentially more serious way: I'll treat the various credence functions as if they obey the rules of probability. As noted above, I don't think this is generally right for maximally rational credence functions. But I don't think that the ways probabilistic coherence gets used below spoils the point of the examples, which is just to illustrate one possible way in which a certain kind "level-connecting" principle (one which connects rational beliefs to beliefs about the rationality) may allow for the sort of akrasia required by ITM.

A. Acute Akrasia: Elena, Logic, and Extra-Strength Drugs.

Elena is certain that she's taken an anti-reliability drug that messes with people's performance on truth-functional reasoning about complex social situations (99% of the time, it makes valid arguments seem invalid, and vice-versa; 1% of the time, people assess the arguments correctly). Elena is given the argument from P1 and P2 to C, and asked whether it's valid or invalid. Elena thinks about all this, and rationally adopts .01 credence in V (the claim that the argument is valid). This is what ITM would recommend: the rational credence in V on just Elena's first-order evidence is very high, and treating an agent relevantly similar to Elena's forming this high credence as an indicator, in light of the information about how she was drugged, would give us .01 probability for V.

Now suppose that Elena accepts ITM. So she sees her low credence in V as rational only if the argument is in fact valid, and as way too low if it's in fact invalid. Since she's highly confident that it's invalid, she must be severely akratic. Let's see how her credences fare vis-à-vis NRR:

Given our stipulations about the case, Elena can be certain that one of two situations obtains:

(V) The argument is valid (and the rational credence for V is .01).
(I) The argument is invalid (and the rational credence for V is .99).

Elena has .01 credence in (V) and .99 credence in (I).

Let's use cr for Elena's credence function, pr_V for the credence function that's ideal if (V) obtains, and pr_I for the credence function that's ideal if (I) obtains. Probabilistic coherence gives us:

1. $cr(V) = cr(V \mid \text{pr}_V \text{ is ideal}) \cdot cr(\text{pr}_V \text{ is ideal}) + cr(V \mid \text{pr}_I \text{ is ideal}) \cdot cr(\text{pr}_I \text{ is ideal})$.

Elena's credences in V and I fix the right sides of both summands, so:

2. $cr(V) = cr(V \mid \text{pr}_V \text{ is ideal}) \cdot .01 + cr(V \mid \text{pr}_I \text{ is ideal}) \cdot .99$.

NRR says that the left sides of the summands should be rewritable as follows:

3. $cr(V) = \text{pr}_V(V \mid \text{pr}_V \text{ is ideal}) \cdot .01 + \text{pr}_I(V \mid \text{pr}_I \text{ is ideal}) \cdot .99$.

Since Elena knows that V is true if pr_V is ideal, and that V is false if pr_I is ideal, we get:

4. $cr(V) = 1 \cdot .01 + 0 \cdot .99$.

So cr(V) = .01, and Elena's extremely akratic state still complies with NRR.

B. Moderate Akrasia: Disagreeing about the Weather

Fatima and Gus are very reliable meteorologists, with access the same meteorological evidence, which rationalizes .8 credence in the claim that it will rain tomorrow (R). They have extensive, equally good, track records of predicting rain. Fatima studies the data and arrives at .8 credence in R, then learns that Gus has arrived at .2. For a bit more simplification, let us suppose that the Epistemology Oracle lets them know that one of them (she of course won't say which) has arrived at the credence that's rational on their meteorological evidence. And let us suppose that ITM prescribes a .5 credence in R for Fatima in this situation, and that Fatima in fact adopts .5 credence in R.

How will Fatima, who accepts ITM, see her credence? She'll see that her .5 credence is rational only if she's the one who assessed the meteorological evidence correctly. But she thinks that it's equally likely that Gus was the one who assessed the meteorological evidence correctly, in which case her .5 credence is irrationally high. Let us suppose that ITM would put the rational credence for Fatima in that case at .25 .[37] So Fatima is as confident as not that her credence is significantly too high; this is an example of moderate akrasia.

Applying NRR, we can see that Fatima should recognize two possibilities, and think them equally likely:

(F) Fatima assessed the meteorological evidence correctly (i.e. it rationalizes .8 credence in rain).
(G) Gus assessed the meteorological evidence correctly (i.e., it rationalizes .2 credence in rain).

Let's use pr_F for the credence function that is ideal if (F) obtains, and pr_G for the credence function that is ideal if (G) obtains. Probabilistic coherence gives us:

1. $cr(R) = cr(R \mid pr_F \text{ is ideal}) \cdot cr(pr_F \text{ is ideal}) + cr(R \mid pr_G \text{ is ideal}) \cdot cr(pr_G \text{ is ideal})$.

Fatima's credences in (F) and (G) fix the right sides of both summands, so:

2. $cr(R) = cr(R \mid pr_F \text{ is ideal}) \cdot .5 + cr(R \mid pr_G \text{ is ideal}) \cdot .5$.

NRR says that the left sides of the summands should be rewritable as follows:

---

[37] Why not .2? ITM will assess the reliability of an agent relevantly similar to Fatima reaching .2 credence in R. But I take it that this assessment should take into account not only Gus's reaching .2, but the agent's former self having reached .8. While it's not clear what precise credence ITM would recommend in this case (it would depend on evidence about just how reliable Fatima and Gus were at assessing meteorological evidence), it's quite plausible that it would be somewhat higher than .2, but much closer to .2 than .8. For present purposes, however, the exact number will not matter, since it will not affect whether Fatima's credences satisfy NRR.

3. $cr(R) = pr_F(R \mid pr_F \text{ is ideal}) \cdot .5 + pr_G(R \mid pr_G \text{ is ideal}) \cdot .5.$

Now it seems that in Fatima's situation, conditional on (F) obtaining, it's rational to think rain .8 probable, and conditional on (G) obtaining, it's rational to think rain .2 probable. These conditional credences should be reflected in whichever credence function is ideal in Fatima's situation. So we get:

4. $cr(R) = .8 \cdot .5 + .2 \cdot .5$.

So $cr(R) = .5$, and Fatima's moderately akratic state complies with NRR.

References

Adler, J. E. (2002), *Belief's Own Ethics* (Cambridge, MA: MIT Press).

Barnett, Z. (ms.), "Disagreement and Belief Dependence: Showing When and How the Numbers Count".

Berger, D. (ms.), *Rational and Reasonable Beliefs - Two dimensions of Epistemic Criticizability and an Alternative to Conciliationism*, Senior Honors Thesis, Brown University.

Christensen, D. (2007a), "Epistemology of Disagreement: The Good News", Philosophical Review 116: 187–217.

---. (2007b), "Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals," Oxford Studies in Epistemology 2: 3-31.

---. (2010), "Higher-Order Evidence," *Philosophy and Phenomenological Research* 81: 185-215.

---. (2011), "Disagreement, Question-Begging and Epistemic Self-Criticism," *Philosophers' Imprint* 11, no. 6.

---. (2014), "Conciliation, Uniqueness and Rational Toxicity," *Noûs* Early View (DOI 10.1111/nous.12077).

Coates, A. (2012), "Rational Epistemic Akrasia," *American Philosophical Quarterly* 49: 113-124.

Cohen, S. (2013), "A Defense of the (Almost) Equal Weight View," in D. Christensen and J. Lackey, eds., *The Epistemology of Disagreement: New Essays* (Oxford: Oxford University Press).

Elga, A. (2007), "Reflection and Disagreement," *Noûs* 41: 478-502.

---. (2013), "The puzzle of the unmarked clock and the new rational reflection principle," *Philosophical Studies* 164(1): 127-139

Enoch, D. (2010), "Not Just a Truthometer: Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement," *Mind* 119: 953-997.

Feldman, R. and T. Warfield, eds. (2010), *Disagreement* (Oxford: Oxford University Press).

Greco, D. (2014), "A Puzzle about Epistemic Akrasia," *Philosophical Studies* 167: 201–219.

Horowitz, S. (2013), "Epistemic Akrasia," *Noûs* online first (DOI: 10.1111/nous.12026).

---. (ms), "Predictably Misleading Evidence".

Kahneman, D. (2011), *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux).

Kelly, T. (2005), "The Epistemic Significance of Disagreement," *Oxford Studies in Epistemology* 1: 167-96.

---. (2010), "Peer Disagreement and Higher-Order Evidence," in Feldman and Warfield.

---. (2013), "Disagreement and the Burdens of Judgment," in D. Christensen and J. Lackey, eds., *The Epistemology of Disagreement: New Essays* (Oxford: Oxford University Press).

Kornblith, H. (2010), "Belief in the Face of Controversy," in Feldman and Warfield.

Lackey, J. (2010), "A Justificationist View of Disagreement's Epistemic Significance", in A. Haddock, A. Millar, and D. Pritchard (eds.), *Social Epistemology* (Oxford: Oxford University Press).

---. (2013), "Disagreement and Belief Dependence: Why Numbers Matter," in D. Christensen and J. Lackey, eds., *The Epistemology of Disagreement: New Essays* (Oxford: Oxford University Press).

Lasonen-Aarnio, M. (2014), "Higher-Order Evidence and the Limits of Defeat," *Philosophy and Phenomenological Research* 88: 314-345.

---. (2015), "New Rational Reflection and Internalism about Rationality," *Oxford Studies in Epistemology 5*: 145-171.

Schechter, J. (2013), "Rational Self-Doubt and the Failure of Closure," *Philosophical Studies* 163: 428–452.

Schoenfield, M. (2014), "Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief," *Noûs* 48: 193-218

---. (2014), "A Dilemma for Calibrationism," *Philosophy and Phenomenological Research* Early View (DOI: 10.1111/phpr.12125).

---. (forthcoming), "Bridging Rationality and Accuracy," *Journal of Philosophy*.

Sliwa, P and S. Horowitz (2015), "Respecting *all* the evidence," *Philosophical Studies* online first (DOI 10.1007/s11098-015-0446-9).

Titelbaum, M. (2015), "Rationality's Fixed Point (Or: In Defense of Right Reason)", *Oxford Studies in Epistemology* 5: 253-294.

Weatherson, B. (ms.), "Do Judgments Screen Evidence?"

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

Williamson, T. (2011), "Improbable knowing," in T. Dougherty, ed., *Evidentialism and its Discontents* (Oxford: Oxford University Press).

---. (2014), "Very improbable knowing," *Erkenntnis* 79: 971-999.