[Final version in *Philosophical Perspectives*: *Epistemology* (2010): 121-140. Please feel free to email me for a pdf of the published version: david.christensen@brown.edu.]

Rational Reflection[1]

David Christensen, Brown University

1. A Natural Idea

Intuitively, there seems to be a connection between what one is rational to believe, and what one is rational to believe one is rational to believe. If we think of beliefs in a categorical, rather than a graded, way, a plausible thought is that rationally believing that P is incompatible with rationally believing that one's belief that P is not rational. Putting the thought in terms of justification, the idea is that (justified) higher-level doubts about the justification of one's belief that P can defeat one's justification for believing P.

If we think of belief in a graded way, however, the connection between the two levels of belief appears in a somewhat different light. There seem to be plenty of cases where one's rationally having a certain credence in a proposition is compatible with one's rationally doubting that it is the rational credence for an agent in one's epistemic situation to have. Suppose, for example, that Ava is considering the possibility that (D) the next U.S. President will be a Democrat. She gives D some particular credence, say .7; this reflects a great deal of her general knowledge, her feel for public opinion, her knowledge of possible candidates, etc. But given the possibility that her credence is affected by wishful thinking, protective pessimism, or just failure to focus on and perfectly integrate an unruly mass of evidence, Ava very much doubts that her credence is exactly what her evidence supports. This seems only natural; indeed, eminently reasonable. And in particular, her judgment about the rationality of her credence doesn't seem to undermine the rationality of her maintaining that credence.[2]

---

1 This paper began in a reading group session on Williamson (ms. 2007) led by Josh Schechter, and subsequent discussions with Schechter were crucial to the paper's development. I was also helped in thinking about these issues by attending Roger White's seminar on Skepticism, by subsequent correspondence with White, and by very helpful conversations with Jonathan Vogel. Earlier versions were presented at the Formal-Traditional Epistemology Workshop at the University of Stirling, and at Epistemology: The Third Brazil Conference at PUCRS in Porto Alegre; thanks to audiences at both talks for stimulating questions, and to Christopher Clarke, Branden Fitelson, Carl Hoefer and Ralph Wedgwood for subsequent comments and discussion.

2 One might protest that, given the messiness of her evidence, Ava shouldn't have a sharp credence, such as .7, but should have some spread-out credal state. We needn't settle that issue

Part of the reason that Ava's meta-level judgment does not undermine the rationality of her credence in D is that it does not lopsidedly suggest that her credence deviates from what's rational in one particular direction. In contrast, suppose that Brayden also gives .7 credence to D. But then Brayden, a staunch Republican, comes to believe, on the basis of ample psychological evidence about himself, that he almost always becomes irrationally confident of unpleasant possibilities, and never underestimates them. When he discovers this, he comes to believe that his credence in D is likely to be irrationally high. Here, it would seem that Brayden is not rational in maintaining his .7 credence in D in the face of his meta-level belief. He should become less confident of D. Once he does that, he may then be in a position like Ava's: he will have doubts that his credence is exactly the right one, but those doubts will not undermine the rationality of his new credence.

Ava's and Brayden's cases seem to throw some light on general rational connections between first-order graded beliefs and higher-order beliefs about the rationality of those first-order beliefs. In certain cases, rational second-order credences do seem to put constraints on rational first-order credences. Is there a general principle that encodes this connection, a graded-belief analogue to the sort of principles that are attractive in the case of categorical belief?

A natural candidate is what might be called Rational Reflection. If we use Cr for an agent's credences and Pr for the credences that would be maximally rational for someone in that agent's epistemic situation, we get something like the following principle, which I'll call Rational Reflection:

RatRef: Cr(A / Pr(A)=n) = n

It seems plausible, at least initially, that a rational agent's credences should obey RatRef. It basically says that the agent would take the maximally rational response to her epistemic situation as a certain sort of guide to the truth. To violate it would be as if to say, for example:

---

here (though the protest is not obviously correct--see Elga (2010) and White (forthcoming); though see also Sturgeon (forthcoming). The point would remain even if Ava had some particular spread-out credal state: given the messiness of the evidence, isn't it likely that some other (perhaps differently spread-out) state would be better supported by the evidence? In the remainder, I'll represent credences by sharp values; I believe that this will not affect the arguments. I will also ignore issues raised by agents giving credence to infinitely many propositions.

"My own credence in rain, supposing that the maximally rational credence in rain *in my very evidential situation* is .9, is .6!"

RatRef also seems consonant with our differing judgments about Ava and Brayden. The principle does not require that one be confident that one's credence in A is maximally rational. It permits a case where one has .7 credence in A, but thinks that there is some chance that the maximally rational credence for A in one's situation is .8, and an equal chance that it's .6. But it does prohibit, e.g., a case where one has .7 credence in A, but is certain that the maximally rational credence for A in one's situation is .6.[3]

Despite the initial attraction, however, the principle leads to puzzling results in certain cases. In the next section, I'll describe such a case. In subsequent sections, I'll look at some different possible reactions to the case. Thinking through some of these possibilities will, I hope, begin to throw some light on the complexities involved with inter-level connections among rational beliefs.

2. A Puzzling Case[4]

Suppose Chloe is looking at an unmarked clock a few feet away with just a minute hand. The minute hand moves in discrete jumps, so it's always at some exact minute. As she's seeing it, the hand is pointed somewhere in the upper-middle of the lower-right quadrant. Chloe is considering various propositions about the hand's position, such as:

P19: The hand is at 19 minutes after the hour;
P20: The hand is at 20 minutes after the hour;
P21: The hand is at 21 minutes after the hour;
etc.

[3] The way RatRef constrains credences will be developed in more detail below.

[4] This central case in this section is a simplified variant of an example Timothy Williamson (ms., 2007) has used to argue that knowing that P is fully compatible with its being incredibly improbable on one's evidence that one knows that P. His argument is in various ways specific to the knowledge case, and to his particular notion of epistemic probability. But the central idea behind his example is well-suited to raising questions about other inter-level epistemic connections.

How should her credence be distributed over these propositions?

First, it seems clear that she should not be fully confident of any one hand position. Suppose it looks most likely to be a bit below where the 4 would be, so if she had to bet on one position, she'd bet on P21. But if she's reasonable, it wouldn't really surprise her much to learn that the hand was actually on 20, or 22. In fact, she shouldn't be shocked to learn that it was on 19 or 23. More distant positions, of course, would be more surprising, and she may be extremely sure it's not lower than 16 or higher than 25. So it seems that her confidence should be distributed over the possible hand positions in a tight, roughly bell-shaped curve, with the peak at 21. A description of these rational credences would assign to each Pn the credence she should have that the minute hand is at minute n.[5] Part of that distribution, let's suppose, looks like this:

... P19: .1; P20: .2; P21: .3; P22: .2; P23 .1; ...

Given that this sort of pattern is rational in Chloe's epistemic situation, it seems that similar patterns (but with the peak of the curve moved around) will be rational in situations like Chloe's, but where the visual evidence provided by the minute hand's position is different. Given information about the clock-room setup, and the way Chloe's visual system works, it seems that one could in principle describe what credence distribution would be rational for Chloe in each of the 60 evidential situations she might find herself in, when confronted by the clock. Such a description would look like a big 60 X 60 chart showing, for each hand position, Chloe's rational credence in each proposition about hand position. Let's call this Chloe's Chart.

Suppose that Chloe is believing maximally rationally: the hand is in fact at 21, and Chloe has exactly the beliefs—excerpted above—that the Chart prescribes for her particular evidential situation. So far, this seems unproblematic. Chloe may not be sure she has exactly the rationally correct credences, but her doubts on this topic seem like Ava's: they don't undermine the rationality of the credences she has.

[5] Again, this involves a simplifying assumption; this time that there is exactly one maximally rational credence-distribution for Chloe's evidential situation. This assumption is certainly controversial (see, e.g., White (2005); Kelly (forthcoming)). At this point, I do not believe that it will affect the substance of the argument below, though I'm not sure. In any case, I'll continue working with this assumption.

But now suppose we share with Chloe our Chart describing the epistemic facts that apply to her. And suppose she has very good reason to trust us, and she does so; for now, let's assume for simplicity that she becomes certain that the Chart is correct. It seems that trouble ensues, as follows: Chloe is .3 confident that P21 (the hand is exactly at 21), and .7 confident that it's somewhere else. But she also thinks that it's rational for her to have .3 confidence that P21 only if she's actually in the evidential situation that would be produced by the hand being at 21. And she believes that the rational credence for her to have that P21 is less than .3 if she's in any of the other evidential situations she might be in. Since she thinks she's probably in one of those other situations, it seems that she should think that .3 is probably too high a credence for her to have that P21, and certainly not too low. But that seems to suggest that her credence that P21 should be lower than .3. Her situation, in other words, is now intuitively more like Brayden's than like Ava's. Yet the claim that she should lower her confidence that P21 below .3 looks inconsistent with our assumption that the Chart correctly describes her situation.

This statement of the problem is deliberately loose and informal, with no reference to any specific principle connecting higher-level and lower-level credences. And I think that the intuitive problem arises quite independently of how we formulate an inter-level connection principle. But looking at the problem through the lens of RatRef will provide a more sharply focused version of the difficulty.

To see this, note that in the clock example, Chloe will end up violating RatRef (at least if she's coherent). This is because Cr(P21) must be equal to the (weighted) average of her conditional credences Cr(P21 / Pr(P21)=n), for all values of n where Pr(P21)=n gets positive credence. Given Chloe's certainty about the Chart, RatRef ensures that the highest conditional credence feeding into the average is .3; this is the term corresponding to the situation where the hand is at 21. All the rest of the terms are lower than .3, so the weighted average must come out below .3. This is inconsistent with Chloe's unconditional value for Cr(P21) being .3. So once Chloe becomes certain that the Chart applies to her, her credences are coherent only if she violates RatRef.

In fact, it seems that RatRef is straightforwardly inconsistent with the claim that a coherent agent in Chloe's situation can be rationally certain that the Chart applies to her. After all, given certainty about the Chart, Chloe will be certain that that

a) Pr(P21)=.3 iff P21

But certainty of this guarantees (modulo coherence) that:

b) Cr(P21 / Pr(P21)=.3) = Cr(P21 / P21)

Clearly, the right-hand side of (b) must be 1. But RatRef says that the left hand side should be .3.[6]

What should we make of this example? Williamson takes his somewhat analogous example not as presenting a puzzle, but as showing how knowledge is quite independent of certain sorts of higher-level probabilities. We could take an analogous tack, deny RatRef (along, presumably, with any similar inter-level principle), and hold that Chloe's .3 credence in P21 remains perfectly rational, even after she learns about the Chart. Another option is to deny that the Chart describes the credences it would be rational for a person in Chloe's situation to have (even absent her knowledge of the Chart). Still another option is to hold that that the Chart is possibly correct, but only if unknown—in other words, the Chart may apply to someone in Chloe's situation only if she doesn't know about it (or, more precisely, only if she isn't certain—or even very confident—of what it says). In the next three sections, I'd like to look at these options.

3. The Split-Level Strategy

Clearly, RatRef is difficult to square with the intuitively plausible idea that one might be in a situation like Chloe's, and be certain that the Chart applied. And despite the history of people offering Reflection-style constraints on rationality, there's also a history of people finding problems with them. So one might well wonder whether the whole puzzle flows from a superficially appealing, but fundamentally misguided, formal principle.

The motivation for RatRef derives from our sense that one's *beliefs about* what would be rational to believe in one's situation put constraints on what it is *actually* rational to believe in one's situation. As many people have noticed, in thinking about categorical belief, there seems

[6] Strictly, the inconsistency arises when some non-0 credence is given to P21 and Pr(P21)=.3.

to be something wrong with believing P while simultaneously believing that it's not rational to believe P in one's situation. Of course, as William Alston (1980) has emphasized, it's important to be careful here. We would not want, for instance, to make it a precondition of justifiedly believing P that one (justifiedly) believe that one was justified in believing P. That way lie worries about regress, and about denying justified beliefs to unsophisticated agents. But such problems do not seem to be entailed by rejecting the rationality of simultaneously (1) believing that P and (2) believing that one is not justified or rational in believing that P. And as Richard Feldman (2005) has argued recently, there is a strong intuitive case to be made for rejecting the rationality of this pair of attitudes.[7]

Putting the issue in terms of categorical belief simplifies things in a way. The evidence that your belief that P is not justified is evidence that you've erred in a particular direction—by believing P when you shouldn't have. But as we saw above, the situation is more complex when one thinks in terms of degrees of credence. If I have some degree of credence in P, and doubt that it's the rational one, that alone doesn't tell me if I should worry that I'm too confident, or not confident enough. That's why Ava's situation doesn't seem obviously problematic. It's only when the higher-order information indicates a particular direction for one's expected deviation from rational correctness—as it does in Brayden's case—that the information seems to call for rational repair.

Will this difference undermine the basic intuitions behind the thought that others have defended in the categorical case: that rational higher-order beliefs put constraints on rational lower-order belief? If so, perhaps we could then just revise our initial judgment about Brayden, reject anything like RatRef, and hold that Chloe may simply retain her .3 rational credence in P21, even after she knows about the Chart. In order to explore this possibility, let's begin by considering in more detail an example in which an agent gets evidence suggesting that her credence in a certain proposition is not rationally supported by her evidence.[8]

---

[7] See also Bergmann (2005, 423) and Gibbons (2006, 32). Jonathan Adler (2002) argues for an even stronger thesis: that one literally cannot believe P while simultaneously believing that one's evidence does not support P, at least if one is fully aware.

[8] The example is a slight variant of one given by Adam Elga (ms., 2008).

Hypoxia is a deficiency of oxygen getting to the brain; it commonly affects people such as mountain climbers and pilots who are exposed to the low partial oxygen pressures encountered at high altitudes. In the initial stages of hypoxia, the victim's judgment is compromised, but the victim feels just fine. Naturally, this fact features prominently in the many sources warning pilots about hypoxia:

> The onset of hypoxia is insidious. Rarely does a person recognize that he or she is affected. A deterioration of flying skills-the inability to hold altitudes, entering incorrect radio frequencies-can be an example of the onset of hypoxia. (Shelton, 1999)

> *The most dangerous aspect of hypoxia is that the individual experiencing hypoxia does not and cannot detect the decrement in function and loses the ability for critical judgement.* (Carlson 1998), italics original.

10,000 feet is often mentioned as the altitude at which one should start to worry about hypoxia (though numerous factors can affect a particular individual's susceptibility). With this in mind, consider the following case:

Pilot: You're alone, flying a small plane to a wilderness airstrip. You're considering whether you have enough fuel to make it safely to an airstrip 50 miles further away than your original destination. Checking the relevant factors, you become extremely confident that you do have sufficient fuel—you figure that you've got a fair bit more than the safety-mandated minimum. So you begin your turn toward the more distant strip. But then you notice that your altimeter reads 10,825 feet. You feel completely clear-headed and normal; however you're fully aware of the insidious effects hypoxia can have. Should you trust your recently formed confident judgment about having sufficient fuel, and continue on your path toward the more distant airstrip?

It seems clear to me that you would be grossly irrational if you were to rely on your recent reasoning about having sufficient fuel to make the more distant airstrip. And this is quite independent of whether you are in fact hypoxic. Even if, unbeknownst to you, your altimeter is

malfunctioning, and you're actually well below 10,000 feet, the altimeter reading still gives you clear evidence that you might be hypoxic. Moreover, the obvious practical irrationality of relying on your recent fuel-assessment derives directly from the epistemic irrationality of maintaining your high level of confidence that you have enough fuel for the lengthened flight: if it were epistemically rational for you to believe with undiminished confidence that you had sufficient fuel, you'd be practically rational to act on that confidence.[9]

It's important to note that the possibility of hypoxia, in itself, provides no direct evidence one way or another about whether your plane has enough fuel. The evidence provided by the altimeter seems to undermine the rationality of your maintaining your initial level of confidence only by way of raising doubts about whether that level of confidence is rationally supported by your initial evidence. So this looks like a clear case where rational high confidence in P can be undermined by having reasons to believe that one's confidence in P is not rationally supported by one's (first-order) evidence.[10]

Of course, this isn't yet to arrive at RatRef. But the example serves to support the general claim that rational first-order credences are constrained by one's higher-order credences about which first-order credences would be rational in one's situation. And the constraint seems consistent with what RatRef would require.

So let's now look at some other cases, to see if the principle is a reasonable way of making precise the worry that we have in the pilot case. Suppose that I'm a horse-racing enthusiast who knows a lot about how to predict race outcomes based on various sorts of historical data, information about track conditions, etc. I'm confronting a large mass of data, and wondering how likely it is that (M) Mr. Ed will finish in the money. Thinking as hard and

---

[9] Elga also argues that his pilot cannot rationally remain confident in her judgment after learning she's at risk for hypoxia. A similar conclusion is endorsed for categorical belief by Joshua Schechter (ms., 2010).

[10] One might wonder how this connects to the point made above about *direction* of error. After all, I've given no general reason to think that hypoxia would make one over- as opposed to under-confident in having enough fuel. I'll take this up in more detail below, but for now, let me point out that given your extremely high initial confidence in having enough fuel, the maximally rational confidence cannot be much higher, but it could well be much lower. Intuitively, the import of the altimeter evidence depends not only on how likely your credence is to have deviated in a given direction, but also on how far it might have deviated in that direction.

clearly as I humanly can, I arrive at a .4 credence for M. But I'm worried that I may have made a mistake—after all, the information I have is very complicated, and I know I'm not perfect at thinking about complicated matters. Given my desire to win money, I find myself saying out loud,

"If only the Epistemology Oracle would tell me what the rational credence in M is, given all this data about the race!"[11] Lo and behold, she appears, and tells me,

"The maximally rational credence in M, given your data about the race, is .7!" Then she vanishes, leaving me to clean up the epistemic mess. What should I believe about the horserace now?

Intuitively, it seems clear that, insofar as I have reason to trust the Oracle,[12] I should raise my credence in M. If I'm absolutely certain that what the Oracle tells me is true, this seems to speak for raising my credence to .7. After all, what could justify my having lower credence, except evidence? And I'm certain that the evidence makes .7 credence rational. And even if we reject the idea that I could be rationally absolutely certain of the Oracle's pronouncement, it seems that insofar as I'm rational in being very close to certain that the Oracle is correct, that tells in favor of raising my confidence in M so it's very close to .7.

How does this intuitive verdict fit with RatRef? I can't simply apply it by taking the Oracle's pronouncement to show directly that Pr(M)=.7. The Oracle gave me the maximally rational credence for M, given the evidence I had when I called out to her; but I now have another bit of evidence: the Oracle's pronouncement. And the Pr(M) that figures in RatRef is supposed to take into account all of my evidence. Still, it seems clear that if the maximally rational credence in M given my evidence up to the pronouncement is .7, that adding the Oracle's pronouncement to my evidence will not (significantly) move the maximally rational

---

[11] I learned of the existence of the Epistemology Oracle from Roger White (see, e.g., White (2005).

[12] One might question here whether I should believe the Oracle, especially if my original evidence did in fact support .4 credence in M. I think it's intuitively clear that I should believe the Oracle in versions of the case where I have excellent evidence for her reliability, and strong evidence of my own fallibility. And that holds even when the Oracle in fact speaks falsely, and my initial assessment of the horse racing evidence was correct (for discussion of this type of issue, see Christensen (2010)).

credence in M away from .7.[13] So I should be very highly confident that Pr(M)is (at least very close to) .7. And if that's right, then RatRef will underwrite our intuitive judgment that I should move my credence in M so that it's very close to .7.

Now the case as I described it is especially simple, since I'm virtually certain what the rational credence in M for me is. So let's suppose things are a little more complicated, and the Oracle says:

"Your thinking about the race is thoroughly bollixed up! The way you've thought this through, you might as well have picked your credence using a random number generator. In fact, your epistemic performance on this task has been so miserable that Oracle Guild Rules prohibit me from telling you the rational credence. But I can tell you this. The maximally rational credence in M, given your information about the race, is between .6 and .8."

In this case (assuming once more that I have extremely high rational confidence in the Oracle's veracity), I think it's again intuitively clear that I must raise my credence in M dramatically. And again, RatRef can help deliver this result.

Let's look first at what I should believe about Pr(M), after the Oracle's pronouncement. I'm extremely confident that the maximally rational credence given the evidence I had before the pronouncement was between .6 and .8. It seems clear, again, that adding the Oracle's pronouncement to this evidence would not move the maximally rational credence for M away from that range. So I should distribute my credence about the value of Pr(M) (at least almost completely) among values within the .6 - .8 range. And once I've done that, RatRef will ensure that (virtually) all my conditional credences in M given values for Pr(M) will be within this range. And that will ensure that my overall credence for M is within that range.[14]

---

[13] This assumes that I don't have some strange belief about the Oracle which makes her pronouncement bear on my other beliefs in some way other than by way of my taking her pronouncement as true—e.g., a strange belief to the effect that she's more likely to answer my entreaties when I'm enquiring about false propositions. But the case need not involve any such strange beliefs, and the most natural versions of the case do not.

[14] One might think that this was the sort of case in which the point-value model of credences assumed in this paper is inappropriate. Dealing with the issues raised by rejecting the point-value model of credences is beyond this paper's scope. But it's worth noting that even if one were to hold that my credal state in this situation should be located between .6 and .8 in some spread-out way, it would still be true that my credal state with respect to M was constrained by my credences in propositions about the rational credence in M. My believing the Oracle's

So it seems to me that RatRef nicely captures an attractive way of generalizing the intuitive result in the first horserace case to cases where one isn't sure what the rational credence is. It has the advantage of allowing this sort of uncertainty, while still putting some constraints on belief.

Something like RatRef also seems to mesh nicely with our verdict in the hypoxia case. There it looked as though, insofar as we think that you should become much less confident about having enough fuel when you notice the altimeter reading, our judgment is sensitive not only to the direction of possible deviation from rationality suggested by the altimeter, but also to the degree of possible deviation. We saw that although you may have had just as much a chance of becoming overconfident as underconfident, the fact that the possibilities of overconfidence involved much bigger deviations from the rational credence was part of what convinced us that, in taking account of this, you should become less confident. This is consistent with the basic idea that your credence in F (that you have enough fuel to safely go to the further airstrip) should be the weighted average of your credences in F conditional on various possible values of Pr(F) (the maximally rational credence in F, given your evidence), where those conditional credences obey RatRef. RatRef ensures that credences in F conditional on very low values for Pr(F) will themselves be very low—much lower than your initial credence in F. When these are averaged with the credences conditional on higher values of Pr(F)—which cannot be much higher than your initial high credence in F—the average should come out much lower than your initial credence in F.

Now the pilot case is not one in which you are provided with direct information about rational credences given your evidence. But it seems that the source of your concern on reading the altimeter is precisely that it gives you reason to give significant credence to possibilities in which the rational credence for F given your evidence is quite different from the high credence you had before noticing the altimeter. So RatRef seems like a natural candidate for explaining why it is that you would be irrational to maintain your high confidence in F.

Thinking about these cases, then, seems to do two things. First, and most obviously, it strongly suggests that the Split-Level strategy is not an attractive option in general for theorizing

---

pronouncement—like Chloe's believing in the Chart—requires a dramatic change in my first-order credence.

about rational credences: evidence that supports certain sorts of doubts about the rationality of our credences can exert rational pressure to modify those credences. This makes it difficult to see how Chloe could simply maintain the confidence she had in P21, despite believing what the Chart tells her. To do so would seem, at least on the surface, to be analogous to my retaining my low degree of credence in Mr. Ed finishing in the money, even after believing what the Oracle tells me, or to your maintaining undiminished confidence in having enough fuel, even after noticing the altimeter reading and believing what you do about hypoxia.

The second thing that thinking about these cases suggests is that the sorts of modifications in lower-level credence called for by high-level doubts can be captured by something like RatRef. Of course, none of these considerations constitutes a solid case for RatRef. We've seen only that it seems to mesh with intuitively attractive verdicts in a few cases. But even if we end up rejecting RatRef, the cases suggest that some constraint along roughly similar lines applies, at least in many cases. And this is enough, it seems to me, to suggest that we look at other possible approaches to Chloe's predicament.

4. Rejecting the Values in the Chart

Another possible explanation for the puzzlement generated by Chloe's case is that the claims made by the Chart are in some way intrinsically defective. As we've seen, Chloe's certainty that the Chart applies to her is straightforwardly inconsistent (modulo coherence) with RatRef. Perhaps this indicates that, even aside from issues brought on by Chloe's becoming certain that the Chart applies to her, the Chart misdescribes what credences it would be rational for someone in Chloe's type of situation to have.

Of course, rejecting the Chart raises the question of what the correct Chart would look like. What credences would be rational for someone with Chloe's visual capacities who was looking at an unmarked clock? The Chart specifies, for each evidential situation such a person could be in—which we are taking to be correlated with the position of the clock's hand—how her credence should be distributed among propositions about the hand's position. Its salient structural features, which seem to generate the puzzle, are first, that the various positions around the clock are treated symmetrically, and second, that in each evidential situation, the Chart distributes credence so that (a) the true proposition gets more credence than the competing

propositions, and (b) that the true proposition gets less than full credence. So let us consider whether these features are independently suspect.

Clearly, the symmetry feature is not fully realistic, as the case was described. I take it that an ordinary person looking at an unmarked clock face in an ordinary room will be more confident about the hand position if it looks to be right at the top or bottom (and perhaps even if it looks to be at the 15 or 45 minute positions). In part, this is because in an ordinary situation, there will be various clues helping to pick out "up" and "down". And in fact, the whole example presupposes that Chloe has a pretty good idea of which way is up, since otherwise she'd have no idea what minute the hand was pointing to. But it also seems clear that relaxing the symmetry to the point required by realism would not eliminate the problem. Suppose we altered the Chart so that in the evidential situations produced by the hand being at 0 or 30, the agent's credence was much more concentrated in the true proposition. It seems clear that this would have no tendency to ameliorate the problem Chloe gets into when she's in looking at the hand at 21. So I take it that the symmetry is a harmless, and eliminable, simplification.

Let us then consider features (a) and (b) as they apply to Chloe's situation. One might also take (a)—the assumption that all competing propositions get less credence than the true one—to be unrealistic. Presumably this is not because the evidence supports some competing proposition more than the true one. But perhaps it will be suggested, e.g., that for an agent in Chloe's situation, P20 and P22 should get equal credence with P21.

The main thing to notice about this suggestion is that, assuming that more distant possibilities—say, P19 and P23—get some smaller amount of credence, the problem remains undiminished. For Chloe will still be giving some credence to possibilities in which the rational credence for P21 is lower than hers, but she won't be giving any credence to possibilities in which the rational credence for P21 is higher than hers. So the conflict with RatRef remains, as does the informal version of the problem deriving from Chloe having reason to estimate the rational credence in P21 as lower than her present credence.

To avoid this result, it would seem that one would have to hold—quite unrealistically, it seems to me—that the Chart should, e.g., assign equal 1/3 credences to P20-P22 and assign *no* credence to any other possibilities. In that case, the problem with Chloe's credence in P21 would not surface, assuming that the Chart treated P20 and P22 like it treated P21. (Each of the possibilities countenanced by Chloe would be ones in which the rational credence for P21 was

1/3, so with respect to P21, she'd be in RatReflective equilibrium.) But even in this case, there would be trouble with Chloe's credence that P20. For while two of the possibilities Chloe countenances would be ones in which the Chart gave P20 1/3 credence, the other possibility (P22) would be one in which the Chart-mandated rational credence for P20 was 0. So Chloe would again be faced with a credence—her 1/3 credence that P20—that was higher than her expected rational credence for that proposition.

It seems, then, that if there's an objectionable feature of the Chart that's responsible for Chloe's problem, it's (b)—that it assigns less than full credence to the true proposition. Now at first blush, it seems obvious that any correct Chart would have this feature. And I think that the intuitive obviousness persists at least through the second blush, especially given that we can easily vary the example so as to involve more finely-divided dials. So is there anything to be said in favor of faulting the Chart for saying that Chloe should be less than maximally confident in P21 when she has the visual evidence she has in her current situation?

It seems to me that there actually is something at least odd about this feature of the Chart. We can see this by contrasting the Chart's prescriptions for Chloe with other cases in which it seems correct to prescribe non-extreme credences. Consider, for example, a doctor who has (and should have) .8 credence that her patient has hepatitis, based on her list of the patient's symptoms and on statistics she knows which say that 80% of patients with these symptoms have hepatitis. In this sort of case, the fact that the doctor should not be maximally confident that her patient has hepatitis seems to derive from the fact that in evidential situations *exactly like the present one*, the patient has hepatitis only 80% of the time. The doctor's rational credence is limited, as are most credences most people have most of the time, by the fact that her evidence doesn't discriminate perfectly between cases in which the relevant proposition is true, and ones where it's false.

Chloe's Chart, however, presents a different picture. Notice first that it prescribes a distinct set of credences for each of the 60 evidential situations Chloe could be in. The plausibility of this is predicated on the natural thought that Chloe's visual experience will be different in each situation (albeit very slightly different in adjacent situations), and the thought that differences in visual experience will produce differences in the credences those experiences make rational. But once we hold that Chloe should respond differently in distinct evidential situations, one might well ask why it is, given that the evidential situation Chloe is in only occurs

when P21 is true, that don't we hold that Chloe should have full credence that P21 when she's in that situation? There's a sense in which Chloe's experience *does* discriminate between cases where P21 is true and ones where it isn't. In this respect, the limitation on Chloe's rational credence is different from the limitation on the doctor's. Thus there might seem to be a certain kind of tension embodied in the Chart's prescriptions: On the one hand, it in a sense holds Chloe responsible for reacting differentially to adjacent evidential situations. On the other, it implicitly recognizes that adjacent situations are indistinguishable to Chloe, at least in the following specific sense: when she's in a given situation, she can't rationally be fully confident that she's not in an adjacent one.

One possible reaction to this point would be to eliminate the tension in favor of holding that Chloe should be absolutely certain that P21. One might defend this initially unintuitive move by saying that the Chart's credences represent *ideally rational* beliefs, and that while we would not expect an actual human to live up to these ideals, that shouldn't undermine their status as ideally rational.

But it seems to me that our initial intuitive rejection of the rationality of Chloe's being certain is in the end correct. To be rationally certain that P21, Chloe would have to be certain that she wasn't mistaking the P20 visual experience for the P21 experience.[15] Of course, it may be claimed that a perfectly rational agent would be immune from such cognitive errors. But this claim, it seems to me, is beside the point. For even if it is a fact that Chloe is cognitively perfect, and never misinterprets her experiences, she has no reason to be certain of that fact. Even if, say, she would in fact always pick the correct hand position if she were forced to pick just one, she has no grounds for being certain that this is so. So it seems that she cannot absolutely dismiss the possibility that she's made the sort of mistake in question. And to the extent that she can't dismiss that possibility, she must countenance the possibility that P21 is false.[16]

---

[15] This is not the doubt that P20 could, for some physiological reason, result in the visual experience canonically associated with P21—the assumption of the Chart is that this doesn't occur. So we may assume that Chloe is fully confident of the Chart's assumption that her visual experiences may be typed by the hand position. Rather, the doubt in question is that she could be misinterpreting the experience itself, which is, of course, extremely similar to the visual experience she'd have in adjacent situations.

[16] Thanks to Jonathan Vogel for help on this point.

So while one might argue that a Chart mandating absolute certainty of the correct hand position in each evidential situation did capture some aspect or dimension of rationality, it does not seem that such a chart would actually describe ideally rational credences. Our initial intuition—that extreme confidence in the true hand position would be irrational—was correct. And if that's right, then it's hard to resist accepting that our initial Chart (or some Chart of the same structure) is the best description of rational credences for someone in Chloe's position—at least before she's shown the Chart.

5. Does Chloe's Seeing the Chart Falsify It?

Of course, it's one thing to say that the Chart correctly describes rational credences for someone in Chloe's initial situation, and another to say that the Chart continues to correctly describe her rational credences once she sees, and fully believes, what the Chart says. For it's possible that showing Chloe the Chart changes her evidential situation (or, perhaps more broadly, her epistemic situation).

It might at first seem odd that if someone is already believing completely rationally, just giving her *accurate* information about what is rational to believe in her evidential situation could *change* what it's rational for her to believe. But there are cases where this happens in what strikes me as a non-puzzling way. Consider a variant of the example where I'm a bettor interpreting a complex body of horseracing evidence, and arrive at a .4 credence in a M. In this variant, when I call out to the Epistemology Oracle, she tells me: "The rational credence in M, given your horseracing evidence, is at least .4." It seems to me that in this case (insofar as I rationally believe what the Oracle says) I should move my credence in M upwards. After all, given the complexity of the evidence, I should initially be quite doubtful that I've assimilated it all perfectly. Given what the Oracle tells me, .4 either too low or exactly right. Surely the possibility that it's too low is still very much alive, after the Oracle's pronouncement. So I should become more confident in M.[17]

[17] One might object that, depending on what I believe about the Oracle's habits, the Oracle's pronouncement might actually also eliminate the possibility that my credence is too high. This might occur if, for example, I was certain that the Oracle only gave pronouncements to people who had reacted perfectly to their original evidence. But as long as I'm not certain that such a story is true, the Oracle's pronouncement will leave open the possibility that my credence is too low.

Note that this would hold even if .4 happened to be the correct credence for me to have (absent the Oracular pronouncement). So it turns out that simply giving me accurate information about what the rational credence in M is, given my current evidence, can change what credence it's rational for me to have. And on reflection, this is not so mysterious. In effect, the Oracle's pronouncement gives me new evidence—some information about possible errors I might have made. Intuitively, the Oracle is telling me that I haven't overestimated the likelihood of M, but she leaves open the possibility that I have underestimated it.

Now there's some reason to think that Chloe's situation is different from my situation with the Oracle. The Oracle gives me information that is true; but intuitively, the information is skewed. If she told me that in my situation the rational credence was at most .4 and at least .4, then the credence it would be rational for me to have in M would not change. Chloe's chart, by contrast, is not obviously skewed in this way. Nevertheless, the horserace case shows that there's nothing terribly strange about a particular credence being rational on evidence E, but not rational once E is supplemented by some true information about what belief it's rational to have on E. And, as in the horserace case, the information the agent gets in Chloe's case eliminates the possibility that she's made an error in one direction, but not in the other.

Does this, then, provide a non-puzzling resolution to our puzzle? I think that, on closer inspection, Chloe's case turns out to have some more interesting consequences. To see why, let's consider a version of the case that does not turn on Chloe being given new information after she forms her credence in P21.

In the original version of the case, our imaginary Chart was based on our knowledge of the testing setup, and of Chloe's visual system. But this is the sort of knowledge that Chloe might well have herself, even before entering the room with the clock. And our assuming that Chloe has this sort of knowledge of the setup and of her visual system does not seem to change the intuitions behind the Chart: if Chloe has this sort of knowledge, it still seems that, were we to draw up a Chart for her, it would have the general shape of our original Chart. But now, instead of imagining us giving Chloe a copy of the Chart, let's just imagine Chloe asking herself, before entering the room with the clock, "What does the Chart describing the rational credences for me in this sort of testing situation look like?" Suppose first that, highly rational agent that she is, Chloe becomes virtually certain of what is in fact the correct answer to this question—that is, she thinks up the same 60 X 60 Chart we would have drawn up for her, and becomes confident that it

correctly describes her situation. Then she looks at the clock, and forms the Chart-mandated credence in P21. It seems that instability ensues, in exactly the same way it did when we handed her the Chart.

What should we make of this? One might want to hold that by merely thinking through the question of what is rational for her to believe, Chloe has changed what's rational for her to believe. But this seems not to solve the problem. After all, she may have asked herself, "What's rational for someone like me to believe, *on thorough reflection*, when looking at the clock?" And it does not seem that the true answer to this question should be significantly different from our original Chart in its general structure. As we've seen, it wouldn't make sense for Chloe to conclude that she can be rationally certain of the correct hand-position. And it's obvious that it wouldn't make sense for her, while looking right at the clock, to be completely agnostic, giving all hand-positions equal credence. And the remaining, sensible, possibilities seem to have the troubling structure of the original Chart: highest credence in the true proposition, tapering off across adjacent cases. So whatever changes thorough reflection might or might not induce in the exact values entered in on the Chart, it seems that if Chloe thinks accurately about the question of what her considered credences should be, and becomes highly confident of the correct answer to the question, it will cause instability.[18]

It seems, then, that Chloe's case presents a puzzle that goes well beyond the fact that one can sometimes change what a person is rational to believe by informing them of some truth about what they (currently) are rational to believe. As a first approximation, the puzzle is that the epistemic truth for Chloe is unknowable, in this sense: Chloe apparently cannot rationally believe in the accuracy of the general Chart which in fact correctly describes the rational credences for someone in her situation. More precisely, Chloe cannot place high credence in the accuracy of this Chart without violating some rational ideal.

If that is correct, the puzzle deepens as follows. We may suppose that Chloe has ample *evidence* supporting the correctness of the relevant Chart. So once Chloe asks herself what the

---

[18] It is worth noting that this sort of problem does not seem to arise in the most recent horserace case. There, it seems that if I reflect thoroughly on what credence I ought to have, I will not be led into destabilizing conclusions. And while the Oracle's pronouncement does change what it's rational for me to believe, it does not self-undermine in the way Chloe's chart seems to: I can coherently believe that what the Oracle tells me applies to me, even after I've accepted the pronouncement.

correct Chart is for someone like her looking at the clock, it's not obvious that there's any way she can emerge epistemically unscathed. There seems to be no coherent way for her to combine (a) having rational credences about what credences are rational for her to have, and (b) having rational credences about the position of the clock hand, while (c) respecting the rational constraints the former place on the latter. If Chloe is reflective, it's not clear that there is any set of credences she can adopt that will satisfy all the relevant rational ideals. So it does not seem that the puzzle can be dissolved simply by noting that getting information about what's rational to believe in one's epistemic situation may change what it's rational to believe.[19]

6. Conclusion

What, then, should we conclude about Chloe's predicament? We arrived at it by beginning with a natural thought: that part of being a rational believer involves taking into account the possibility that one has made epistemic errors; and that a central part of rationally taking this sort of possibility into account involves adjusting one's degrees of confidence in lower-order propositions in response to higher-order doubts. If we think of belief in graded terms, RatRef is a simple, natural way of expressing the relationship that seems to hold between one's ground-level credences and one's higher-level credences about what beliefs are rational in one's situation. As we've seen, however, obeying RatRef seems, in certain cases, to require agents to violate other rational ideals.

Of course, it is entirely possible that RatRef, even if it is on the right track, needs refinement, and that the correct refinement will solve the problem with Chloe's case while leaving the refined principle capable of accounting for the pilot and horserace cases. One way of refining the principle would be to restrict its application to cases not involving certain sorts of "inadmissible"

[19] One might point out that this example involves the idealization that Chloe can become rationally highly confident in the one correct Chart. Now I don't see any barrier in principle to an agent's being sufficiently informed about epistemology and her visual system that she could arrive at the correct Chart. And even given that the example involves idealization, I don't see that being cognizant of this sort of idealization reduces what's puzzling about the case. (Also, it seems likely that the idealization could be relaxed without dissolving the puzzle. It seems plausible that Chloe could at least be fairly opinionated about the Chart—i.e., concentrate her credence in a small batch of closely bunched charts. And that should suffice for undermining in many cases.)

information about Pr(A), where one's account of admissibility would count Chloe's Chart as inadmissible.[20]

Now any such restriction on RatRef would have to be motivated, and it's worth noting that the motivation for restricting RatRef might be particularly problematic. Principles such as RatRef, Lewis's Principal Principle, or van Fraassen's Reflection are intuitively designed to take a certain probability function—the inner one—as a kind of "expert". Such principles are likely to have intuitive exceptions when the agent has reason to think that the expert is likely to be misguided in a certain case, or when the agent has some evidence that the expert function hasn't taken into account. But in the case of RatRef, the expert function Pr seems by definition to take into account all of the agent's evidence that bears on the matter in question, and to do so in the maximally rational manner. So motivating restrictions will be more tricky. That said, it's certainly possible that some more sophisticated relative of RatRef might allow us to avoid puzzlement in Chloe's case.

Another (possibly complementary) angle that might be explored involves self-reference. There's a sense in which Chloe's Chart—at least in the original case where we hand the Chart to her after she forms her credence in P21—seems self-referential. In order to produce instability, it must be interpreted as giving the maximally rational credences for an agent who has reflected on the Chart's own values. In the later case, where Chloe is given in advance the evidence on which the Chart was based, and comes to the Chart's conclusions about rational credences on her own, the self-reference is less clear. But one might still be suspicious that some problematic sort of self-reference lurks beneath the surface. Still, it's not obvious how one would avoid this sort of reflexivity in thinking about what one should believe. It seems that my being rational sometimes involves my reflection on whether my beliefs are those best supported by my evidence, and my subsequently bringing my first- and higher-order beliefs into alignment. But to do that, it seems that I must ask myself something like, "How strongly should I believe P, on thorough reflection, given my current evidential situation?"

All that said, it seems clear that the ideal outcome of thinking about cases like Chloe's would be our finding a way of accommodating the intuitions behind RatRef while avoiding the difficulties we've been examining. But if we can't find any, there is an alternative to escaping

[20] The possibility of exploring this line, analogous to some treatments of Lewis's Principal Principle, was suggested to me by Branden Fitelson and Carl Hoefer.

the puzzle that's worth mentioning. We might end up acknowledging that, in certain cases, agents such as Chloe have no choice but to violate some perfectly respectable epistemic ideal.

While this may seem puzzling, I don't think it would make Chloe's case paradoxical, or even unique. I would argue that somewhat similar dilemmas confront agents—even ideal thinkers—who contemplate logical truths: since such agents cannot be certain that they've reasoned ideally, they should have less than full confidence in logical truths, and hence violate probabilistic coherence.[21] Of course, one may reject taking probabilistic coherence as a rational ideal. But other sorts of cases, not involving probabilistic coherence, seem to fit the same pattern. One sort involves the unfortunate agents in the literature on epistemic paradoxes who are given good reason to believe something of the form: (I'll believe that P) iff ~P. There, it seems plausible to think that agents will violate some epistemic ideal no matter what they end up believing. And I think that the same sort of phenomenon occurs in cases involving ordinary agents who get evidence—based, say, on drugs, or sleep-deprivation, or on the disagreement of highly skilled thinkers—that their ordinary opinions on contingent matters are based on mistakes in their thinking. If such agents must, at least to some extent, bracket or put aside the challenged reasoning in assessing the possibility that that reasoning is flawed, then they too are likely to end up being forced to violate epistemic ideals.[22]

If we take this sort of view of Chloe's situation, it's worth noting that this does not commit us one way or another on the question of whether there is one best way for her to resolve her epistemic dilemma. Still less does it commit us to saying that the epistemically best response for Chloe will have her satisfying RatRef. We would only be holding that Chloe is under rational pressure to avoid a certain kind of incoherence between her first-order and higher-order credences, and that this problematic incoherence can be understood in terms of violating RatRef.

At this point, I would certainly not claim that RatRef is correct—a great deal more thought needs to be given to the general topic of inter-level connections among rational credences. But I would submit that the fact that RatRef leads to puzzles in Chloe-like situations doesn't by itself show that it is not on the right track.

---

[21] See (Christensen 2007) for more on this.
[22] This is argued in detail in (Christensen 2010).

If RatRef, or something in the same neighborhood, does turn out to be correct, it may extend the range of cases in which we can see that agents are precluded from satisfying every rational ideal. But the extended range of cases would still be unified in one way. The complexities they involve arise from a dimension of belief-management that's crucial to any agent who must confront the possibility of her own fallibility: the agent's critical reflection on her own beliefs.

References

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

Alston, William P. (1980), "Level Confusions in Epistemology," *Midwest Studies in Philosophy V: Epistemology*: 135-150.

Bergmann, Michael (2005), "Defeaters and Higher-Level Requirements," *The Philosophical Quarterly* 55: 419-436.

Carlson, Robert W. (1998), "Supplemental Oxygen for the General Aviation Pilot (What you don't know might kill you)," accessed 7/20/10 at http://www.dr-amy.com/rich/oxygen/.

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

---. (2010), "Higher-Order Evidence," *Philosophy and Phenomenological Research* 81.1.

Elga, Adam (ms., 2008), "Lucky to be Rational," presented at the 2008 Bellingham Summer Philosophy Conference.

---. (2010), "Subjective Probabilities Should be Sharp," *Philosophers' Imprint, 10(5), 2010.*

Feldman, Richard (2005), "Respecting the Evidence," *Philosophical Perspectives 19: Epistemology*: 95-119.

Gibbons, John (2006), "Access Externalism," *Mind* 115: 19–39.

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

Schechter, Joshua (ms., 2010), "Rational Self-Doubt and the Failure of Closure."

Shelton, Joe (1999), "Hypoxia: Coming Down for Air," *Plane and Pilot* June 1, 1999, accessed 7/20/10 at http://www.encyclopedia.com/doc/1P3-41650300.html.

Sturgeon, S. (forthcoming), "Confidence & Coarse-Grained Attitudes," *Oxford Studies in Epistemology.*

White, R. (2005), "Epistemic Permissiveness," *Philosophical Perspectives* 19: 445 - 59.

---. (forthcoming), "Evidential Symmetry and Mushy Credence," *Oxford Studies in Epistemology*.

Williamson, Timothy, (ms., 2007), "Improbable Knowing [notes]," http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0014/1319/Orielho.pdf.