[Final pre-publication version. Published version in *Philosophical Studies*, at https://doi.org/10.1007/s11098-020-01536-6, or email me for a copy: david.christensen@brown.edu.]

Akratic (Epistemic) Modesty[1]

Abstract: Theories of epistemic rationality that take disagreement (or other higher-order evidence) seriously tend to be "modest" in a certain sense: they say that there are circumstances in which it is rational to doubt their correctness. Modest views have been criticized on the grounds that they undermine themselves—they're self-defeating. The standard Self-Defeat Objections depend on principles forbidding epistemically akratic beliefs; but there are good reasons to doubt these principles—even New Rational Reflection, which was designed to allow for certain special cases that are intuitively akratic. On the other hand, if we construct a Self-Defeat Objection without relying on anti-akratic principles, modest principles turn out not to undermine themselves. In the end, modesty should not be seen as a defect in a theory of rational belief.

Introduction

Epistemologists are fond of formulating, and defending, principles characterizing rational belief.[2] The principles are supposed to be general, covering beliefs on all sorts of topics—including, of course, beliefs about what principles characterize rational belief. This reflexive aspect of epistemic theorizing has led to some perplexity lately, in discussions of epistemic principles which entail that agents in some circumstances are rationally required to doubt the correctness of those very principles. Let's call such principles 'modest'.[3]

Much of the perplexity has arisen in the literature on disagreement, in discussions of Conciliatory views about rational belief. On these views, the disagreement of apparent epistemic peers can dramatically lower the credence it's rational to have on the controversial topic. Critics have objected to Conciliationism along roughly the following lines: Given the controversy over Conciliationism among qualified epistemologists, adherents of the view cannot believe their own view—they must have strong doubts that believing Conciliatorily is rational. So if they continue to abide by Conciliationism, they'll be believing in a way they strongly suspect to be irrational: they'll be epistemically akratic. Since this cannot be a rational way of believing, Conciliationism must be defective. Let's call this the 'Self-Defeat Objection'.[4]

---

[1] Many thanks to Nomy Arpaly, Jennifer Carr, Branden Fitelson, Sophie Horowitz, Chris Meacham, Ram Neta, Richard Pettigrew, Josh Schechter, Mattias Skipper, the participants in my seminar at Brown, participants in the Harvard Workshop on Bounded Rationality, and Julia Staffel, my commentator at the Workshop, for very helpful comments and questions. Thanks also to an anonymous referee for *Philosophical Studies*. And special thanks to Zach Barnett for both positive suggestions and perceptive pushback.

[2] The notion of rationality I'm interested in here is a distinctively epistemic one. If there are principles governing which beliefs it's pragmatically rational to have (or, perhaps more plausibly, to try to get oneself to have), they are a separate matter. On the notion under discussion, it may be rational for a father to believe his daughter committed an awful crime, even if he realizes that having that belief will destroy him emotionally.

[3] This is one use of 'modest'; there are other uses in the literature.

[4] A number of different versions of the Self-Defeat Objection have been offered: See Elga (2010), Weatherson (2013), Weintraub, (2013), Decker (2014), and Reining (2016). I'll concentrate here on the version that strikes me as most serious, which will be described in detail below.

Defenders of modest principles have pointed out that this sort of self-defeat applies way beyond Conciliationist views of disagreement. It applies to almost all current views on disagreement—since they entail that at least disagreement by large numbers of epistemic superiors requires at least some loss of confidence. And it applies beyond the disagreement issue entirely, to many accounts which require agents to take "higher-order evidence" seriously—for example, accounts which mandate diminished confidence in complicated claims when agents know they're in situations where fatigue, bias, or drugs make them very likely to misevaluate the claims in question.[5] But while this point suggests that there must be *something* wrong with the Self-Defeat Objection, or *some* nice way of saving modest principles in the face of the Objection, it doesn't in itself provide insight into how the Self-Defeat Objection can be avoided or accommodated. So defenders of modest principles have offered a range of more specific responses; these generally concede that the Self-Defeat Objection has significant force, but argue that there nonetheless are ways of saving modest principles.

Elga (2010) proposes that Conciliationism be modified to specifically exempt belief in Conciliationism from its own scope. Frances (2010) suggests that Conciliationism shouldn't be believed, but may rationally be employed as a rule of thumb in guiding our believing. Christensen (2013) claims that modest principles are correct, but that the Self-Defeat Objection reveals that they lead to inevitable violations of epistemic ideals—in particular, an ideal forbidding epistemic akrasia. Pittard (2015) argues that the Conciliatory agent faced with the disagreement about Conciliationism must either violate Conciliationism or violate the reasoning method behind it—and since she can't avoid both, she can rationally refuse to conciliate on Conciliationism. Matheson (2014) and Reining (2016) argue, in different ways, for weakenings of Conciliationism to avoid the objection. Ballantyne (2019, ch. 10) considers several responses without defending any of them, but argues that their disjunction is more plausible than the denial of Conciliationism. I won't enter here into criticizing the specific responses (complaints about many of the proposals in the list above can be found in the subsequent papers on the list). But I think it's fair to say that each of them involves significant intuitive costs, or concessions from the proponent of modest epistemic principles.

A few writers have taken a different tack, putting forward a more radical view: that higher-order evidence does not actually affect rational belief. On this sort of view, agents are always rational to believe as their first-order evidence dictates, and information about disagreement, fatigue, bias, or drugs makes no difference. This position in a way avoids the whole problem, in that it automatically precludes Conciliationism and all accounts of rationality that take higher-order evidence seriously. But the intuitive costs are pretty steep.[6]

---

[5] See Frances (2010), Christensen (2009, 2013).

[6] Positions along these lines are offered by Kelly (2005), Lasonen-Aarnio (2014), and Titelbaum (2015); a related view is put forth in Smithies (2019). The intuitive costs of this sort of position have been brought out in many papers in the literature on disagreement and higher-order evidence. For a representative example, consider a pilot who consults her flight instruments, whose readings in fact rationally support, in a complex way, the claim that she has enough fuel to reach an airport more distant than her original destination. She reaches this conclusion, but then is told that she's likely hypoxic (hypoxia often causes people to reach irrational conclusions in complex reasoning, even while feeling clear-headed). Is the pilot rational to maintain her extremely high confidence that she has enough fuel? The views in question seem committed to saying that she is.

It would be nice, then, if a closer look at the Self-Defeat Objection showed that no concessions were needed in the first place. In the following sections, I'll argue for just that conclusion.

1. The structure of the Self-Defeat Objection

It will be helpful to start with a more detailed picture of how the Self-Defeat Objection proceeds. Here is an instance which targets Conciliatory views of disagreement; it takes the form of a *reductio*:

Suppose that Conciliationism is true. Now consider Alma, a Conciliationist who also has views in philosophy of mind. When Alma considers matters directly, it seems clear to her that Physicalism is true (for the sake of the example, we may suppose that the direct arguments do, in fact, support Physicalism). But she encounters Bento, a confident dualist whom she has good reason to consider an epistemic peer in philosophy—someone who's as likely to reach correct philosophical conclusions as she is. So she lowers her credence in Physicalism—say, from .9 to .5 (the exact numbers don't matter, but let's stipulate that this is the rational credence for her to have, according to Conciliationism). Then Bento mentions that he rejects Conciliationism, too, in favor of a Steadfast view on which facts about other people's opinions on philosophical issues have no bearing on the credence it's rational to have about those issues. So Alma lowers her credence in Conciliationism—say, to .5 (again, the exact number doesn't matter)—and raises her credence in the Steadfast view. Now it seems that Alma cannot think of her own .5 credence in Physicalism as rational. If Conciliationism is true, her middling credence is just right. But if the Steadfast view is right, the rational credence for her is much higher. So her expectation of what the rational credence is for her in Physicalism will be well above .5. If rationality requires that her credence in Physicalism reflect her beliefs about what credence is rational for her, it would seem that she now has to have above .5 credence in Physicalism. But by hypothesis, Conciliationism requires her to have .5 credence in Physicalism. If it requires two different credences in the same proposition, Conciliationism is defective as a theory of rational belief.

The effect can be more pronounced in other scenarios. If Alma reasonably considers Bento to be more skilled at epistemology, or if we add more Steadfast-theorists to the mix, Conciliationism would require Alma to end up with an even lower credence in Conciliationism. So Alma would be required to become even more confident that her .5 credence in Physicalism was irrational. And again, if Alma formed a credence in Physicalism that matched her expectation of the most rational credence, it would be very far from the credence Conciliationism calls rational.[7]

In examining the workings of this argument, it is important to notice what the argument does not show. It does not actually show that Conciliationism is impossible to conform to consistently. Alma clearly could continue to hold her conciliated .5 credence in Physicalism, even while being quite confident that conciliated credences are irrational. It's just that she would then not be 'practicing what she preached' epistemically—she would be epistemically akratic. So the Self-Defeat Objection rests squarely on the claim that this cannot be rational. Given that assumption,

[7] This formulation of the problem draws on Elga (2010) and Christensen (2013). Note that this version of the problem is not dependent on any contingent facts about the present distribution of opinion on Conciliationism. The objection is that if an account of rationality would, in some epistemic situation, require incompatible or irrational belief-states, it must be defective.

modest epistemic principles, which require agents in certain circumstances to become akratic, must be incorrect.[8]

Let us then turn to take a closer look at this crucial element in the Self-Defeat Objection.[9]

2. Intuitive Epistemic Akrasia

What is the relationship between (1) rational beliefs about the world, and (2) rational beliefs about which beliefs about the world are rational?[10] As Arpaly (2000) points out, it clearly doesn't follow from a person's believing a certain belief to be rational that the belief is rational for the person to hold: all sorts of people confidently believe their irrational beliefs to be rational. And most reject the principle that for a belief to be rational, the believer has to have some meta-belief in the belief's rationality. But the claim the Self-Defeat Objection depends on is considerably weaker. The idea (put roughly, in terms of categorical belief) is something like this: it cannot be rational to simultaneously have a certain belief and believe that belief to be irrational. Putting the idea in terms of credences is a bit trickier (it's not obviously irrational, for example, for me to have .5 credence in some claim when I think that the rational credence for me to have in the claim is either .45 or .55, but where I see these possibilities as equally likely). But the worry would clearly arise for a case where someone had very high credence in some claim while also being highly confident that only a very low credence in that claim was rational in their situation. These ideas can be made precise in a number of different ways, but let us put that off for now, and refer to the intuitively troubling phenomenon as *intuitive epistemic akrasia*.

Epistemic akrasia is widely seen as irrational—even obviously so.[11] But there are a number of different kinds of cases which should give the defender of enkratic requirements—that is, requirements forbidding akratic pairs of beliefs—pause. Let us begin by looking at two of them; these will challenge some, but not all, enkratic requirements.[12]

---

[8] Several discussions of the self-defeat problem take this facet of the argument explicitly into account. See Littlejohn (2012, 2014, 2020), Christensen (2013), Pittard (2015), Matheson (2015), and Titelbaum (2015).

[9] I have presented the Self-Defeat Objection in the standard way, as a challenge based on doubts about whether modest principles give correct accounts of *rational* belief. There is, of course, the question of whether a similar Self-Defeat Objection might be based directly on doubts about whether following modest principles yields *accurate* beliefs. I will discuss this issue in detail in sections 5 and 6 below.

[10] Here and below, I'll use "the world" as shorthand for whatever is the subject-matter of the first-order beliefs in question. Of course, one might dispute about whether beliefs about a priori matters are about "the world," or one might argue that beliefs about rationality are beliefs about "the world". I hope I can use this shorthand without expressing commitment to any particular answers to these questions.

[11] See, for example, Adler (2002), Bergmann (2005, 423), Gibbons (2006, 32), Smithies (2012), or Littlejohn (2018). Even Arpaly (2000), which argues that when an agent has false beliefs about what's rational, the most rational option is often the akratic one, takes akrasia as violating a rational coherence condition.

[12] One might think that this issue would be moot. Some have thought that Conciliationism would be completely unmotivated without some sort of enkratic principle. Weatherson (2013) argues this explicitly, and the line is implicit in the popular thought that Conciliationism is motivated only if disagreement is evidence that one's original belief is irrational. And the same thought would apply equally to other modest accounts of rational belief.

a. Rational uncertainty about one's evidential situation

Horowitz (2014), adapting a case from Williamson (2011), describes the following scenario: You are to throw a dart at an unmarked dartboard, which is inlaid with an invisible magnetic grid; darts can only land at intersections of the grid lines. The grid is fine enough that you can't rationally be sure, just by looking, of the coordinates where your dart has landed. But wherever it lands, you can rationally be certain that it's within a particular limited area. To fix ideas, suppose that, wherever the dart lands, this area includes the four adjacent points in addition to the point where the dart landed. And suppose that the rational credence for you is equally distributed among these 5 points, and, finally, that you're aware of all these facts.[13]

Now suppose that your dart lands at point p. By assumption, your rational credence that it's at p is .2. More interestingly, your rational credence in "Ring-p" (the claim that the dart is at one of the four points adjacent to p) is .8. But you're aware that a .8 credence in Ring-p is only rational in one evidential situation: the one where the dart is at p. At every other point, the rational credence in Ring-p is much lower. So you are rationally required to have .8 credence in Ring-p, but also to be .8 confident that your credence in Ring-p is irrationally high, and that only a much lower credence would be rational!

Horowitz, who argues that most instances of akrasia involve irrationality, attributes the surprising result to a strange feature of the situation. In the dartboard case, and related cases from Williamson, the agent should expect rational credence and accuracy to come apart: high credence in Ring propositions are rational only when they're false (that is, when the dart is in the center of the donut). And whenever a Ring proposition is true, the rational credence in it is low (.2).

b. Rational uncertainty about one's reliability

Christensen (2016) argues that rationally required akrasia will fall out of any theory of rationality that respects plausible intuitions about cases where agents get powerful evidence that their thinking on a certain subject has been distorted to the point of anti-reliability. Consider a case where Chen gets evidence that she's been dosed with a powerful drug that makes people get the wrong answer in sentential logic problems 90% of the time, even while feeling totally clear-headed. Then she's given a sentential logic problem where the actual correct answer is 'invalid'.

Let us consider how a maximally rational agent would respond to this total evidence. It seems that they would first see the invalidity of the argument (the fact that they have strong evidence that they've been drugged obviously doesn't entail that their sentential-logical reasoning is

---

But this would, I think, be too quick. Treatments of disagreement have been divided, and sometimes ambiguous, between seeing disagreement as evidence that one's initial thinking was *irrational*, and seeing it as evidence that one's initial thinking was *inaccurate*. Elga's (2007) early defense of Conciliationism, for example, was clearly formulated in terms of accuracy, not rationality. See Christensen (2014) for extended discussion, and argument that the conciliatory force of disagreement primarily flows from concerns about accuracy, not rationality. I will discuss accuracy-based versions of the Self-Defeat Objection below.

[13] The exact numerical details are not required for the example, which also works even with non-uniform distribution of credence among the points were the dart (for all you can tell) may have landed. The main assumption is that you cannot be certain which evidential situation you are in.

actually compromised). Next, they would compensate for their expected malfunction, not trusting their own direct assessment of the problem: they would end up with high credence that the argument is actually *valid*. This is simply the result of responding rationally both to the first-order evidence (the logic problem), and then to the evidence of anti-reliability.

Suppose the correct epistemic theory accommodates this intuitive verdict—that is, that the maximally rational belief for agents in situations like Chen's will be inaccurate. And suppose further that Chen herself is rationally confident in the theory's correctness. As we have seen, the theory holds that Chen should have high confidence in 'valid'. But what should she think about the *rationality* of her credence? According to our epistemic theory, which Chen believes, agents who have strong anti-reliability evidence about themselves, and who are looking at *valid* arguments (which she believes she is), are rationally required to have *low* credence in 'valid'. So Chen should be highly confident that her *high* credence in 'valid' is irrational, and that only a much lower credence would be rational!

But, one might ask, if Chen thinks that only a much lower credence in 'valid' would be rational, how can she rationally maintain her high credence? Again, the answer flows from a rationally expected divergence between rationality and accuracy. Our epistemic theory entails that, in cases where agents have strong anti-reliability evidence about themselves, maximally rational agents will be misled. Since Chen can see that this applies to her present case, she can see that her credence being irrational would not indicate its inaccuracy—indeed, quite the opposite is true.

Both of these cases suggest that intuitively akratic states may be the most rational states available, at least in certain odd cases where agents are uncertain about what their evidence is, or about their own reliability. But the explanations of why akrasia makes sense in these cases at least suggest something stronger: that at bottom, there's nothing at all irrational about epistemic akrasia *per se*. In most cases, we should expect rational beliefs and accurate beliefs to go hand-in-hand. And when we should expect this to be the case, a belief's irrationality is an indication that it's inaccurate. So the usual irrationality of epistemically akratic belief-states is derivative: it flows from agents having beliefs they have good reason to believe inaccurate.[14]

However, there is also a way in which these cases still allow for a robust rationally required relation to hold between an agent's first-order beliefs and her beliefs about which beliefs are rational. True, our cases involve agents who are epistemically akratic in some intuitive sense. But there are many ways of formulating enkratic principles. Elga (2013) proposes a New Rational Reflection (NRR) principle, according to which an agent's beliefs about what beliefs are ideally rational in her situation rationally constrain her first-order beliefs. If we let cr_A stand for the credences of a rational agent, and cr_X stand for credences the agent thinks might be rationally ideal in her situation, the principle reads as follows:

NRR: $cr_A(P \mid cr_X \text{ is ideal}) = cr_X(P \mid cr_X \text{ is ideal})$

NRR says that an agent's credence in P, conditional on a certain credence function being ideally rational in the agent's situation, must take the same value that the credence function assigns to P (on

[14] Horowitz (2014) and Christensen (2016) both offer diagnoses along this line.

the condition that it is the rational credence function). More intuitively, we might think of NRR as trying to capture the sense in which ideally rational credence-functions might be thought of as "experts".[15] The principle roughly requires that, to the extent that I believe a certain function to be ideally rational in my present situation, I have those credences as my own.[16] The principle is certainly one way of imposing an enkratic requirement on rational belief.

Interestingly, NRR is explicitly designed to *allow* for the sort of *intuitively* akratic beliefs we've seen in Williamson-style cases like Horowitz's, where agents are uncertain about what evidence they have: such agents do not violate NRR. And Christensen (2016) provides cases suggesting that rational beliefs in cases such as Chen's are also consistent with NRR. Since NRR is clearly a kind of enkratic principle, it's not clear that the sorts of cases we've been considering come anywhere close to vindicating epistemic akrasia in general. If there is an enkratic requirement that has teeth in many situations, perhaps that is enough to give rise to the problem of self-defeat.

In the next section, then, let us turn to examine this principle, and the phenomenon of epistemic akrasia more broadly.

3. NRR and Rational Uncertainty about the Requirements of Rationality

Let us begin with an example closely based on one by Zach Barnett.[17] It involves an agent who has reason to give significant credence to a particular incorrect epistemic principle. Suppose that Dara is a college student who has taken several courses from the epistemologists at his college, who are all smitten by (their understanding of) Hume. Dara's professors espouse a theory of rationality which includes the following principle:

> **Deductive Purism:** Inductive reasoning is not a rational way of supporting beliefs. So, for example, if one wonders whether the sun will rise tomorrow or not, or whether the next bread one eats will be nourishing or poisonous, it's not rational to think either alternative more likely because of what's happened in the past. Inductively-supported beliefs are certainly *accurate* by and large; but rationality is not just about accuracy—it's about support by the right sorts of reasons. Only deductive reasoning can render *rational* support.

Deductive Purism is, of course, false. In fact, the point of the example requires our seeing this: If we're wondering about whether epistemic akrasia can be rational, it won't help to consider cases where agents are *right* about the requirements of rationality, but believe against those requirements—such agents' beliefs will be irrational on any account. The interesting cases will be ones where agents are wrong about what principles govern rational belief. So in thinking about this case, we should not

---

[15] Elga's principle is modeled on the New Principal Principle, which is designed to capture the sense in which objective chance should be taken as an expert.

[16] More precisely, NRR requires rational agents to have the credences that match the ideal function's *conditional* credences, where the condition is that it is the ideal function. For the reasons behind this subtlety, see Elga (2013). I should also note that Elga's principle, being defined in terms of conditional probability, incorporates the assumption that rational credences, and the credences an agent might think are rational, are probabilistically coherent. I think that this assumption should in the end be rejected. But I think that the main points to follow do not turn on this issue.

[17] See Barnett (2020 p. 15 ff). Barnett's example is also aimed at defending the rationality of certain instances of epistemic akrasia.

ourselves give credence to Deductive Purism, but rather examine the implications of an agent's giving credence to it.

And while Deductive Purism is false, it might well be that Dara is rationally required to give it a good deal of credence. Dara's professors confidently and convincingly espouse Deductive Purism. He knows that they are professional experts in epistemology, so he has excellent reason to regard them as reliable sources of information about rationality. Thus it seems in at least some versions of the case, that the most rational attitude for Dara to take is to give high credence to the correctness of the principle.[18]

Dara then begins to wonder about his own beliefs—for example, his belief that the sandwich he packed for lunch will nourish him rather than poison him. A bit disturbed, he asks his favorite professor, "Can't I even rationally believe that my sandwich won't poison me? I mean, otherwise, shouldn't I just throw it away?"

"Oh, Dara—*please* don't throw out your lunch!" his professor smiles. "Look, we never said you couldn't form *accurate* beliefs by reasoning inductively. Of course you can! It's just that inductively supported beliefs are not *rational*. After all, there's no non-circular way to justify induction—remember?"

It seems clear here that even if it's rational for Dara to have high confidence that Deductive Purism is a correct principle of rationality, he is also rational to remain confident that his sandwich will nourish him. He understands perfectly well that on the Deductive Purist account, agents with inductive evidence are faced with the choice of believing accurately but irrationally, or avoiding irrational beliefs by sacrificing accurate beliefs. Since he's clearly in this situation, it seems eminently epistemically rational for him to remain highly confident in the *truth* of the claim that his sandwich will nourish him, even if he's also highly confident that this belief is not *rational*.[19]

How does this case of intuitive akrasia fare vis-à-vis NRR? Interestingly, this case looks different from the ones we've seen so far. According to NRR, Dara's credences, if rational, would satisfy the following equation (I'll use cr_{Dara} for Dara's credences, cr_{DP} for the credences deemed rational on Deductive Purism, and N for the claim that Dara's sandwich will nourish him):

---

[18] I'm putting aside one sort of view here, according to which the ideally rational, or most rational, response to any evidential situation always includes maximal certainty about what the correct rational principles are. On this sort of view, rational beliefs about the principles of rationality are immune to undermining by higher-order evidence. (Views in this neighborhood are defended in Elga (2010), Titelbaum (2015), Littlejohn (2018), Neta (2018), and Smithies (2019).) On this view, Dara would be rational to have complete certainty in the correct theory of rationality (whatever that is), no matter what all the experts said, or, e.g., whether he was presented with powerful evidence that his thinking about the correct theory of rationality had been compromised by judgment-distorting drugs. See Lasonen-Aarnio (2020) for criticism of this sort of view.

[19] One might worry that Dara—or his professors—are not taking Humean skepticism seriously enough. After all, wouldn't it entail not only that one is not rational to believe inductively-supported conclusions, but also that one is not rational to believe *that induction is reliable*—that beliefs in inductively-supported conclusions are *accurate*?

Putting questions of Hume-interpretation aside, we may grant that Deductive Purism does have the implication that one is not *rational* to believe in the reliability of induction (it will at least have this implication if one thinks that the reliability of induction is based on inductive support). But Deductive Purism explicitly does not equate rationality with reliability. So the Deductive Purist may consistently believe that (1) my belief in the reliability of induction is not rational, and (2) induction is reliable. This is just another instance of the kind of akrasia Dara embodies in his sandwich beliefs.

Thanks to an anonymous referee for raising this worry.

cr_{Dara} (N| cr_{DP} is ideal) = cr_{DP} (N| cr_{DP} is ideal).

In other words, Dara's credence in N, on the supposition that Deductive Purism is the right account of rational credence, has to be the same as the credence Deductive Purism would assign to N as rational for an agent in Dara's situation, supposing that Deductive Purism was the correct account of rational credence.

But will this equation hold? An epistemology including Deductive Purism would likely assign a middling credence as the rational credence for Dara to have in N, since Dara's inductive evidence does not count one way or the other. So then the right side of the equation would have a middling value.[20] But if, as we saw, it's rational for Dara to remain highly confident in N overall, the left side of the equation has to have a high value. This is because Dara is rationally highly confident that Deductive Purism is the correct account of rational credence. Given that fact, his overall, unconditional credence in N has to be close to his credence in N on the condition that Deductive Purism is correct. Evidently, the sides of the equation will not match.

What does this tell us? If we accept the intuitive verdicts in the Dara case, it tells us that NRR is not, after all, a correct constraint on rational credence. The sort of relationship it would require to hold between first-order credences and credences about principles of rationality may hold up in certain cases of intuitive akrasia. But it does not hold up overall. And in particular, it does not hold up in certain cases of rational uncertainty about the principles of rationality.

The case also illustrates *why* NRR fails in certain cases. NRR is intended to capture the way in which rationally ideal credences are a kind of expert. It says, essentially, that to the extent one sees a certain set of credences as ideally rational, one expects them to be accurate indicators of the facts. But on some accounts of rationality, such as Deductive Purism, rationality should not be thought of as an expert—at least not in an accuracy-based sense of "expert". Deductive Purism, in its emphasis on respecting the "right sort of reasons" at the expense of accuracy, clearly encompasses the idea that rationality goes beyond accuracy-conducive features of beliefs. So it does not seem at all epistemically irrational for Dara to think, "Well, the *rational* credence for me to have that my sandwich will nourish me may be middling, but surely it's *true* that it will nourish me!"

Of course, Deductive Purism is not offered here as a serious account of rationality. But many serious writers have offered accounts of rational belief that incorporate frankly non-accuracy-aimed features (some of these views only apply to categorical belief, while others would extend to credences as well). Nozick (1993) offers a theory on which the rationality of beliefs depends in part on evidential support, but in part on the expected practical consequences of holding the belief. On Fantl and McGrath's (2002) account, practical stakes for the believer with respect to truth of the relevant proposition help determine the justification of categorical belief. On Nelkin's (2000) account, intended to cope with the lottery paradox, rational categorical beliefs cannot be based on purely statistical evidence, no matter how strongly it indicates truth of the relevant proposition.

---

[20] I should note that different ways of extending Deductive Purism into a full theory of rationality would have different specific results here. On the common Bayesian approach to rational credences into which NRR most naturally fits, rational doxastic attitudes are traditional sharp-valued credences, which are rooted in *a priori* priors, and evolve as evidence accumulates. On this kind of view, Dara's rational credence in N on Deductive Purism would be the rational prior credence N, which would presumably be moderate. We could fill in the example to stipulate that this is the general theory which Dara has been taught. But actually, any extension of Deductive Purism on which the rational credence in N is not very high will suffice to make the point in the text.

Buchak (2014) reaches as similar conclusion from considerations about the way beliefs figure in our blaming practices. According to Marušić (2015), the rationality of certain beliefs—ones about matters that may hinge on our actions—is affected by non-truth-aimed considerations related to our ability to make sincere commitments. Rinard (2017) holds that the rationality of beliefs is determined by the expected value of having them, which, as she acknowledges, can come widely apart from truth-oriented considerations.[21] Gardiner (2018) describes a number of epistemologists who advocate "moral encroachment"—the view that the epistemic justification of a belief can be affected by considerations involving the moral implications of holding the belief. So it's not as if the idea that rationality of beliefs depends on factors that can pull apart from accuracy considerations is in general beyond the pale. That being so, the sort of problem illustrated by Deductive Purism is not an artificial curiosity—it's a serious failing of NRR.

Of course, the point here is not at all to defend any of these theories of rational belief. But it is important to see that NRR is unmotivated, given the existence of conceptions of rational belief that are defended by plenty of serious philosophers. This sort of fact might at first seem not to be a big deal. NRR, one might argue, finds its most natural home in purely evidentialist epistemologies, and the prominent accounts mentioned above depart from evidentialism.[22] So the mere fact that NRR is inconsistent with these accounts can't be a serious problem—after all, plenty of rational principles are straightforwardly inconsistent with other rational principles that have serious supporters.

But given the nature of NRR, it has a problem much more serious than mere inconsistency with other defensible principles. That is because the whole point of NRR is to allow for, and accommodate, rational uncertainty about rationality. So suppose, as we have been, that the right theory of rationality is something one can be rationally uncertain about. In that case, it would seem that the opinions of serious philosophers are exactly the kind of thing that could bear on one's rational credence in the correctness of various views about rationality. Even if one is an evidentialist, one need not hold that rational agents are required to be certain of the falsity of all non-evidentialist theories. This holds especially clearly for evidentialists who are sympathetic to conciliationism: they after all hold that disagreement evidence can undermine rational confidence even in logical and mathematical truths, so it would be pretty strange to then insist that disagreement could not undermine rational confidence in evidentialism!

So the important point is that NRR is not just inconsistent with the *correctness* of certain views about rationality. It yields unacceptable results when agents merely *give credence* to any of a large number of perfectly respectable views of epistemic rationality. NRR, even though its purpose is to

[21] I should note that neither Marušić nor Rinard quite claim to be theorizing about "epistemic rationality." Marušić uses "epistemic rationality" only to apply to beliefs whose truth does not depend on our actions. But he sees the commitment-related factors as crucial to the *only* kind of rationality that applies to beliefs that do depend on our actions. Rinard rejects separating epistemic and pragmatic rationality, and argues that the pragmatic rationality is the only important kind of rationality that applies to belief.

[22] Thanks to an anonymous referee for raising this point.

account for rational uncertainty about the requirements of rationality, actually precludes giving credence to these views.[23, 24]

It seems, then, that cases of uncertainty about rational principles provide more thorough support for embracing epistemic akrasia. Even the sort of limited enkratic requirement encoded in NRR fails to hold. And if that's right, it's not clear that there's any rational pressure to avoid epistemic akrasia at all.

Reflecting on these cases does more than provide intuitive counterexamples to enkratic requirements. It also provides an explanation of why, and where, these requirements should be expected to fail. There is not space here to engage in detail with the various moves made in the literature in defense of various enkratic principles. But I think it is fair to say that many of the defenses lean heavily on the intuitive strangeness of thoughts in the neighborhood of "P, but my believing that P is not rational." (Horowitz (2014), Littlejohn (2018), and Silva (2018) all offer compelling dialogues in which forthrightly akratic subjects sound very silly.) Some then go on to try in various ways to accommodate enkratic requirement: this may entail, say, denying that one can ever be rationally misled about rational requirements[25], or offering expressivist or contextualist analysis of epistemic terms.[26]

The present cases are fully consistent with the usual silliness of frankly akratic subjects, and the irrationality of most akrasia. In fact, they help *explain* it: in most ordinary cases, subjects are rational to expect that rationality and accuracy go hand-in-hand. But this explanation also allows us to see where akrasia can be rational—even rationally required. As Horowitz points out, the sort of problems she identifies as flowing from ordinary akrasia cases do not arise when the akrasia is licensed by expected rationality/accuracy mismatches. And in explaining away the impulse to reject akrasia *per se*, we avoid having to make implausibly strong claims about our epistemic access to the correct requirements of rationality, or embracing controversial instances of expressivism or contextualism.

With all this in mind, let us turn back to reexamine the Self-Defeat Objection to modest epistemic principles.

4. The Self-Defeat Objection, Revisited

If even NRR doesn't hold in cases of rational doubt about principles of rationality, this suggests that the threat presented by the Self-Defeat Objection may be less worrisome than it might have seemed. And indeed this does seem to be the case.

---

[23] One might wonder here: Why did NRR succeed in the dartboard and drug cases? They too, crucially featured foreseen gaps between rationality and accuracy. The reason is that in both of those cases, the different hypotheses about what credences were rational in the subject's situation each encoded information about the subject-matter of the belief. In the dartboard case, the hypotheses about which credence function was rational each entailed what the position of the dart was. In the drug case, the hypotheses about which credence function was rational each entailed whether the argument was valid or not.

[24] For some very different reasons to worry about NRR, see Lasonen-Aarnio (2015).

[25] See Titelbaum (2015), Littlejohn (2018) and Neta (2018).

[26] See Greco (2014) and Salow (2019), respectively.

Let us return to the example which illustrated the objection. We began by supposing for *reductio* that Conciliationism was true; that Alma believed that it was true; and that the credences she formed were rational (that is, that Alma's beliefs reflected appropriate conciliation with apparent peers). Alma also began with high (.9) credence in Physicalism. She then met Bento, who espoused both Dualism and a (false) Steadfast view of rational belief. Let us elaborate a bit on the case. Suppose that Alma's Conciliationism is embedded in an Evidentialist account of epistemic rationality, on which what's rational for someone to believe depends completely on their evidence. Suppose that Bento's Steadfastness is part of a general epistemic view according to which epistemic rationality requires intellectual autonomy of a certain sort. On this view, the attitude that is rational to take toward P is the attitude that seems right given one's own consideration of the direct evidence for and against P. When another person shares one's direct evidence, letting their take on that evidence influence one's own opinion is an unbecoming abdication of intellectual autonomy, and thus compromises the rationality of the affected beliefs. Finally, to make the problem sharper, suppose that Alma takes Bento to be significantly better at philosophy than she is.

Suppose that Alma rationally conciliates with Bento on both propositions: to simplify, suppose she ends up .3/.7 on Physicalism/Dualism and .3/.7 on Evidentialist Conciliationism/Autonomist Steadfast View. As the Self-Defeat Objection would have it, Alma now has a credence in Physicalism that she herself sees as probably irrational. And this is not a case where she thinks that the rational credence is different from hers, but she's not sure in which direction. She thinks her own credence in Physicalism is likely to be irrationally low, and that her original .9 credence is probably more rational.

Of course, if Alma were to raise her credence in Physicalism, then it would no longer be rational, according to Evidentialist Conciliationism. So we get the purportedly problematic result: Evidentialist Conciliationism requires Alma to be epistemically akratic. And she is not only akratic in the intuitive sense; her credences also clearly violate NRR.

The question that now arises is this: is this a problem for Evidentialist Conciliationism? Are the akratic credences—which Evidentialist Conciliationism requires Alma to have—actually (epistemically) irrational? I see no reason to think so. There seems nothing wrong with Alma thinking that her low credence in Physicalism is probably *irrational*. She thinks that rational credences have to be autonomous, and her low credence in Physicalism is informed by taking Bento to be a more reliable philosophical thinker. But autonomy is not accuracy. Since Alma quite reasonably takes Bento to be a more reliable philosophical thinker, she is also rational in expecting her conciliated low credence in Physicalism to be more accurate than her original higher credence would be. Her continued conciliation yields rational beliefs, despite her having been convinced (falsely) that *rationality* would require different (autonomous) credences.

One might point out that on the Autonomous Steadfast view, Alma's conciliated low credence in Physicalism, and her expectation of its accuracy, are irrational. But that doesn't tell against the actual rationality of Alma's credences, since the Autonomous Steadfast view is false. Alma may be rationally required to give the view considerable credence, but that doesn't mean that we should think that its rationality-verdicts are true. The Self-Defeat threat to Evidentialist Conciliationism is supposed to come from its requirement that Alma give credence to competing theories of rationality, not from the truth of competing theories.

But doesn't all this mean that, on Conciliationism, Alma shouldn't try to have rational credences? Well, something close to this is right. Conciliationism, as understood here, doesn't say

anything about trying. It simply describes the conditions under which certain credences are rational for agents. But it does have this consequence: in certain circumstances, you are not rational to have the credences you (rationally) believe to be rational. This is really just a restatement of the claim that there can be rational epistemic akrasia. And we've seen that it's not hard to make sense of this position. It seems that it can be rational to think that, on various accounts of rationality, rational beliefs and accurate beliefs will predictably come apart in certain situations. (Dartboard, Drugs, Deductive Purism and the Autonomous Steadfast View provide examples.) In some situations where one is rational to believe that the truth of "P" and the truth of "Belief in P is rational" come apart, it seems rationally unproblematic—and quite plausibly rationally required—to believe P while doubting the rationality of that very belief. And if that's right, it's not clear that the standard Self-Defeat Objection gets off the ground.

5. Self-Defeat without Enkratic Requirements?

We've so far been seeing the Self-Defeat Objection in the standard way: as hinging on cases where agents are required by Conciliationism to doubt the correctness of Conciliationism (or, similarly, cases where other modest accounts of rational belief say it's rational to doubt their correctness). And we've seen that insofar as there can be rational daylight between beliefs about what beliefs are rational, and beliefs about what beliefs are accurate, the standard Self-Defeat Objection fails. But perhaps there is a different way of raising a worry in the same general neighborhood.

It is tempting not to distinguish too carefully between believing a certain epistemic principle to be a correct account of rationality, and expecting the beliefs it calls rational to be accurate. But maybe we can put this whole issue aside, and raise the problem more directly. Instead of focusing on agents' beliefs about what principles correctly describe *rational* belief, we might instead focus directly on agents' expectations about the *accuracy* of the beliefs a principle calls rational. Just as a Conciliationist might encounter a philosopher who claims that some non-conciliatory epistemic principle is *correct*—that is, that it gives a correct account of rational belief—, she might also encounter another philosopher who directly claims that the beliefs some non-conciliatory principle calls rational are more *accurate*. A version of the Self-Defeat Objection that relied on this sort case would not need to depend on any sort of enkratic requirement.

To fix ideas, let us consider a toy example that makes the relevant issues transparent. Let us again suppose, for possible *reductio*, that Evidentialist Conciliationism (hereafter EC) is the correct account of rational belief. Suppose that Eva, like Alma, is an Evidentialist Conciliationist. She believes that EC gives the correct account of rational belief, and she also believes conciliatorily, and she expects the credences that EC calls rational in her situation to be most accurate. She's also familiar with the current literature in philosophy of mind, and is highly confident—say, .9—in Physicalism.

Then Eva meets Felix, who is also familiar with the current literature in philosophy of mind, and whom Eva rationally takes to be as likely as she is to reach accurate conclusions about philosophical matters. Felix tells her that he's highly confident in Dualism, and only .1 confident in Physicalism. Eva conciliates, bringing her credence in Physicalism down to .5 (let us assume that this is the credence that EC would deem rational in Eva's situation). But then Felix tells her that he also believes that in philosophy, the credences EC calls rational tend not to be all that accurate. It's much

more accurate, he claims, to align one's credences with those recommended in *Philosophically Appropriate Credences* (hereafter *PAC*), a guide that's updated monthly to take account of the current literature. And he hands Eva a free introductory copy. Let us suppose, for the sake of argument, that EC requires Eva to conciliate fully with Felix on this matter too. So post-conciliation, she has equal expectations of accuracy for EC-approved credences and *PAC*-approved credences. Then, thumbing through *PAC*, Eva notices that it mandates .1 credence in Physicalism!

It might seem that trouble will now ensue. It might be thought that, according to EC, Eva should now mix her credences as follows: 50% for the EC-approved credence, and 50% for the *PAC*-approved credence. With respect to Physicalism, this would mean that Eva should mix .5 (which was rational on EC) and .1 (which is mandated by *PAC*). But that would obviously end up being considerably lower than .5 (say, .3). So—according to this line of thought—EC would directly require one credence (.5) in Physicalism, but would indirectly require an incompatible one (.3), just as in the rationality-based version of the Self-Defeat Objection.

But can this be right? Intuitively, this does not seem like what EC, properly understood, would require. In arriving at the .3 credence in Physicalism for Eva, Felix's opinion seems to have gotten double-counted: once as a direct endorsement of Dualism, and then again, as an endorsement of his book that endorses Dualism. Similarly, if Felix says, "Physicalism is not true. And, furthermore, high credences in Physicalism are not accurate!" that should not be an occasion for double conciliation. This brings out a key difference between this new attempt at formulating a Self-Defeat Objection and the ones discussed in the literature. The standard versions involve beliefs about rules of rationality, and beliefs about the world—and as we have seen, these are quite different things. But the new argument involves beliefs about *what beliefs are accurate*, and beliefs about the world—and there is just not much daylight between these things. An expectation about a particular belief's accuracy would seem to be an expectation about the worldly matter which is that belief's topic.

But even if something like this is plausible intuitively, one might still wonder: can a properly-formulated version of EC avoid the intuitively incorrect result? To get some purchase on this question, let us focus on a specific simple version of EC[27]. Suppose it says that the credence in Physicalism that Eva should end up with, after talking with Felix, is determined as follows: it is the rational credence for her that Physicalism is true, *independent* of Physicalism's first-order support from her evidence, but *conditional* on:

a) Eva, a generally reliable thinker, having reached .9 initial credence in Physicalism;
b) Felix and Eva being equally likely to reach accurate beliefs in philosophy;
c) Felix having .1 initial credence in Physicalism; and
d) Felix having high credence in the accuracy of *PAC*, which recommends .1 credence in Physicalism.

To begin with, we might make the natural supposition that Eva rationally believes that Felix guides his own philosophical beliefs by *PAC*. In that case, it's quite clear that items (c) and (d) in the list above will not have significant independent effects on Eva's rational credence for Physicalism. This is clear from simple examples having nothing to do with disagreement: If my friend tells me that the NOAA has accurate weather predictions, and I see that the NOAA predicts rain tomorrow,

[27] This version of EC is similar in spirit to accounts discussed by White (2009), Cohen (2013), Schoenfield (2015), Sliwa and Horowitz (2015), and Christensen (2016).

that will affect my credence in rain. But if I then learn that my friend has looked at the NOAA prediction, and as a result she believes that it'll rain, this should have no significant further effect on my credence in rain. So in this version of the case, there seems to be no problem at all for EC: it does not have the alleged problematic implication (that the rational credence for Eva is .3).

This is enough to show that the accuracy-oriented version of the Self-Defeat Objection does not work like the standard rationality-oriented version. But it might be worth looking at a couple of variant examples.

First, we might alter the case to suppose that Eva rationally believes that Felix didn't get his credence by following *PAC*, but by independent contemplation of the arguments—perhaps Felix doesn't even know what *PAC* says. In that case, the bits of evidence (c) and (d) would not simply duplicate one another. And the credence that would be rational for Eva would presumably be lower than in the previous version of the example. Again, though, that does not seem problematic. First, it's only reasonable that when Eva has two independent pieces of evidence against the accuracy of her original take on the first-order evidence, her credence should be lower than it was in the first version of the case. Second, there is again no implication that Eva should have two different attitudes toward P: EC simply mandates a single lower credence in this case.

Perhaps this can be shown more starkly with a second variant of our example. Suppose that Eva conciliates with Felix, who reached low credence in Physicalism by thinking directly about the arguments. Then she meets Gerd, whom she also has good reason to consider a peer, and Gerd tells her about the accuracy of the *PAC*-recommended credences. As in the first variant case, Eva's credence will end up lower than .5. But again, that seems just right. It's not significantly different from a situation in which Eva met two peers who (independently) reached low credences in Physicalism. This is not a kind of result that Conciliationism has trouble with.[28]

It seems, then, that having peers who disagree about the accuracy of the beliefs EC calls rational does not present a phenomenon that's different in kind from the phenomenon that EC was designed to account for in the first place: having peers who disagree about factual matters in general. If that's right, then it's not clear that modest views such as EC face any particular difficulty of self-defeat.

6. Some Worries

One thing one might question about the previous section's dismissal of the accuracy-based Self-Defeat Objection is whether it treated the objection fairly. One might protest: "You didn't present the case right. Felix said that *PAC*-approved credences were more accurate than EC-approved credences. But then you had Eva *fully comply with just EC* in taking this disagreement into account. That's not true to the spirit of the objection. She should have partly complied with the *PAC!*"

While this protest clearly stems from a correct observation—that Eva complied fully with EC—, I think that given the dialectical situation, it should not worry us. The Self-Defeat Objection is supposed to be a *reductio*. The Objection is not just that, according to some epistemic principles, EC gives incorrect results—that's just the point that there exist incompatible epistemic principles. The Self-Defeat Objection is supposed to show that modest theories are in some way hoist by their

[28] There are interesting and subtle questions about how to understand the relevant notion of independence among peers' beliefs. See Goldman (2001), Lackey (2013) and Barnett (2019a) for discussion.

own petards. In particular, the point is supposed to be precisely that *complying with EC* leads to problems for EC.

The standard Self-Defeat Objection shows, correctly, that *complying with EC* has the consequence that an agent will (1) have a certain credence in P, and also (2) believe that some different credence in P is more rational. If this sort of akrasia were irrational, then EC would have led to irrationality. But as we saw, akrasia can be rational in cases of rationally foreseen divergence between rationality and accuracy. The present argument is supposed to show, without relying on any enkratic requirement, that EC leads to problems. It would do this if EC required two incompatible credences in Eva's situation. But absent some enkratic requirement, it doesn't.

Could the worry be put in a different way, though? According to Felix, the credence in Physicalism that EC calls rational *after taking his beliefs into account* is less accurate than the PAC-recommended credence. So doesn't Eva have to compromise between whatever EC recommends—even after taking Felix's beliefs into account—and Felix's beliefs? And wouldn't that necessarily be lower than the credence EC recommends after taking Felix's beliefs into account?

Again, it does not seem that EC would have to require any such thing. It's of course true that if Felix does not conciliate, he will still end up disagreeing with Eva's conciliated credences. But Conciliationism does not require multiple conciliations with stubborn interlocutors.[29]

To see clearly why EC need not make incompatible recommendations in cases such as Eva's, we might suppose that EC was formulated completely explicitly, as a function from all possible evidential situations to doxastic responses. (For simplicity, and to make EC as strict as possible, let us suppose that for each evidential situation, EC mandates a unique doxastic response as rational. Allowing for the possibility that EC was permissive would not affect this point—if anything, it would make it less likely that EC would end up contradicting itself.) By construction, one of these evidential situations is the very one we've been looking at, in which Eva finds out about Felix's beliefs, and about what the PAC says. So if Eva's credences comply with EC, Eva will have whatever credence in Physicalism that EC prescribes *in that situation*. There's no reason to think that EC would require anything else of Eva—e.g., that it would require Eva to have credence .5 in Physicalism, and also require Eva to have credence .3 in Physicalism. So again, it looks as though the Self-Defeat Objection simply does not get a foothold.[30]

It's important to see here that there's also no reason to think that EC, precisely formulated, couldn't embody a very robust version of Conciliationism. It surely could. In fact, this is just what's illustrated in the worked examples sketched above. And the correct, precisely formulated, full account of rationality might also incorporate other sources of modesty: it might require agents to

---

[29] See Elga (2010) for a nice explanation of this point.

[30] The point here is similar to that made in Field (2000), and shows how there is something right about Elga's claim that any acceptable view on the epistemology of disagreement "must be dogmatic with respect to its own correctness" (2010, 185). EC will not recommend credences (in the sense of calling them rational) that are different from the credences EC recommends. Understood this way, the claim really amounts to the claim that EC (or any account of rationality) not call the same credence both rational and irrational. And there's no reason to suppose that modest accounts of rationality will run afoul of this requirement. But requiring this sort of consistency is quite different from saying that that EC must be dogmatic in the sense of requiring certainty in *the claim that EC is the correct account of rational credence*; or in requiring that, whenever EC outputs, say, .5 as the rational credence for P, it must output 1 for the rational credence in "*credence .5 is the rational credence in P*."

take into account evidence that they're drugged, biased, sleep-deprived, hypoxic, etc., in the intuitively reasonable ways.

Of course, in certain evidential situations, these principles will require an agent to doubt the correctness of those very principles. (For example, if Felix tells Eva that *PAC* gives the correct account of *rational* credences, EC will require Eva to doubt that EC is a correct account of *rationality*.) And this can sometimes lead to akrasia. But as we've seen, this sort of epistemic akrasia need not betray irrationality.

A somewhat different worry concerns the claim made above that there's "not much daylight" between beliefs about what beliefs are accurate, and beliefs about the world—more specifically, that expecting a high credence in P to be accurate is, essentially, to expect P to be true. But is it really some kind of metaphysical or psychological impossibility to have, e.g., a high credence that high credence in P is very accurate, while also having a low credence in P? I suspect that it may well be possible. So let us suppose, at least for the sake of argument, that it is possible. Might this cause problems for the line taken above?

We have seen how EC can require epistemic akrasia—especially when the agent's evidence supports certain strange views about rationality. But suppose someone's evidence supported strange views about *accuracy* (because, say, of testimony from respectable but misguided authorities)? In that case, couldn't there, after all, be some situations where the most *rational* response for an agent (according to EC) involved having a certain credence in P while expecting a different credence in P to be more accurate? And if that is possible, does this reinstate the Self-Defeat Objection after all?

I think it does not. The main point to notice is that even if this sort of case occurred, it would not mean that EC required an agent to have two different credences in the same claim. When we looked at accuracy-relevant evidence in Eva's case, we saw that it had an effect on her rational credence in Physicalism, since she took evidence supporting *the accuracy of low credence in Physicalism* to lower the likelihood of *Physicalism itself*. Now suppose we change the story to include some evidence that would make it rational for Eva to adopt ideas about accuracy which loosened the connection between these two things. If that happened, then when Eva learned of Felix's belief that *PAC*-approved credences were more accurate, this would simply have a smaller effect on her rational credence for Physicalism. There's no reason to think that she'd be required to have two different credences in Physicalism. To put the point another way, however we attenuate the strength of the connection between claims about the accuracy of credences and claims directly about the world, the function that describes the EC-mandated credences for each evidential situation can take that connection into account without mapping any evidential situation onto two different credences.

It is true that the envisioned type of situation would have some strange consequences. So suppose that Eva talks to some people she has good reason to believe are experts about what accuracy is. Conciliating with them, she reaches some odd view about accuracy—and say that this odd view about accuracy makes the accuracy of beliefs in Physicalism, and the truth of Physicalism, come apart. In such a case, it seems that EC might require Eva to have a certain credence in Physicalism while expecting a different credence to be more accurate. And one might wonder: could this really be the most rational attitudes for Eva to end up with? Perhaps the strongest reason for doubting that such attitudes could be the most rational would flow from supposing that this combination of attitudes somehow violated the logic of 'accurate.' I'm not at all sure of this, myself, but let us grant that something like this is true. In that case, although EC would not require Eva to

have two different credences in Physicalism, it would require her to have credences which violated the logic of 'accurate.'

I think that if this did happen, it would not present any new problem for EC. It has been recognized for some time that any theory of rationality that takes higher-order evidence seriously will even end up requiring credences that fail to respect certain logical relations. So, for example, if I'm doing a logic problem, and (correctly) find that it's valid, but then am told that I've ingested a logic-disrupting drug, or that several acknowledged logic experts believe it's not valid, my most rational response may well involve being highly confident of the premises of a valid argument while being much less confident of its conclusion.[31] So even if the most rational response to conciliating with (misleading) evidence from accuracy-experts did involve failing to respect the logic of 'accurate', that would not, I think, present any new difficulty for modest accounts of rationality.

7. Conclusion

If the above thoughts are on the right track, then a worry that many have had about Conciliationism—a worry which applies equally to many other plausible epistemic principles—turns out not to be much of a worry at all. And the retreating and bullet-biting that some defenders of modest principles have felt driven to turns out not to be necessary.

We need not weaken Conciliationism so that it delivers merely *pro tanto* reasons for belief: a fully formulated Conciliatory account of rational belief may well be thought of as giving necessary and sufficient conditions for (propositional, epistemic) rationality.

We need not see the modest aspect of certain epistemic principles as forcing agents to violate an enkratic epistemic ideal: Modest principles do sometimes require agents to embrace akrasia, but the required akrasia does not violate a rational requirement.

We also need not hold that Conciliationists must stubbornly remain confident in the correctness of Conciliationism, while conciliating left and right on other philosophical matters. They may rationally doubt that Conciliationism gives the correct account of rational belief. But they may still take rational account of their own epistemic fallibility. So when—given the disagreement of thinkers they respect (or the fact that they themselves are likely biased, or hypoxic, or drugged in some reason-distorting way)—a credence different from their initial credence should be expected to be more accurate, they may rationally adopt that credence.

We may also vindicate the line of thought mentioned at the outset. As noted, the Self-Defeat Objection, if it were sound, would eliminate not only Conciliationism, but almost all other accounts of disagreement, including Kelly's (2010) Total Evidence View, Lackey's (2010) Justificationist View, and even Kelly's (2005) original Right-Reasons View, which allowed that disagreements with epistemic superiors could require revisions in belief. It would also eliminate other modest principles which would require modifications in an agent's credence in response to evidence that she was hypoxic, drugged, sleep-deprived, and so on. We can now get beyond thinking "there must be *some* defect in an argument that has these consequences," and see what the defect is.

Finally, we may also avoid biting the bullet that some critics of epistemic modesty have offered us. We need not hold that when agents are confronted with evidence that certain of their

[31] See, e.g., Christensen (2007, 2014).

beliefs are based on thinking that has been compromised, they are rational in simply brushing off this evidence and maintaining full confidence in their original beliefs.

We will not, however, avoid one feature of Conciliationism that some might find embarrassing. Given the current philosophical climate, and given the undoubted intelligence, competence, talents, etc., of certain critics of Conciliationism, it's very plausible that Conciliationism has the consequence that no one is rational to be highly confident in Conciliationism.[32] This consequence does have an ironic sound to it. But it really shouldn't worry the philosopher sympathetic to Conciliationism. Conciliationism does, in present circumstances, say that no one is rationally confident in Conciliationism—but it will say the same about Physicalism, or Four-Dimentionalism, or Internalism (of whatever sort), or any of the other isms that highly competent philosophers disagree about. Part of the motivation for Conciliationism is the realization that the degree of expert disagreement we see in philosophy shows that, over all, we are not highly reliable at reaching correct philosophical conclusions. But even on Conciliationism, Conciliationism doesn't have a special problem with belief-worthiness.

One might well ask, of a Conciliationist-leaning philosopher, how she can go around defending a view she is not highly confident in. This is an interesting question, though it of course applies to all the views our Conciliationist philosopher defends, not just to her Conciliationism. I won't get into answering this question here.[33] But the Conciliationist might point out that the alternative would be for philosophers to go around being highly confident in the truth of the controversial views they defend. Is this an attractive option?

Given the variety of mutually incompatible views that are defended on most controversial questions, it should be clear on anyone's view that most of the (reasonably detailed) views defended today by intelligent, honest, diligent, and highly qualified philosophers are false. A philosopher could, of course, acknowledge this fact while remaining highly confident of his own views—perhaps embracing a sort of Epistemic Exceptionalism. But it seems to me that our Conciliationist philosopher might see taking this sort of stance as not without its own embarrassing implications.

In any event, we need not settle here all the difficult questions about which modest (or immodest) principles of rationality are correct. It is enough that we can now clear the field to pursue these questions freely, without worrying that all modest epistemic principles will somehow self-destruct.

---

[32] Some (e.g. Decker 2014) have found this sort of contingent self-undermining highly problematic, even if it doesn't show Conciliationism false.

[33] One interesting approach involves describing an attitude other than belief that enquirers can rationally take toward the hypotheses they support. For development of this line, see Goldberg (2013), Fleisher (2018), and Barnett (2019b).

References

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

Arpaly, N. (2000), "On Acting Rationally against One's Best Judgment," *Ethics* 110: 488-513.

Ballantyne, N. (2019), *Knowing our Limits* (Oxford University Press).

Barnett, Z. (2019a), "Belief dependence: How do the numbers count?" *Philosophical Studies* 176: 297–319.

---. (2019b), "Philosophy without Belief," *Mind* 128: 109-138.

---. (2020), "Rational Moral Ignorance," *Philosophy and Phenomenological Research* online first: https://doi.org/10.1111/phpr.12684.

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

Buchak, L. (2014), "Belief, Credence and Norms," *Philosophical Studies* 169: 285-311.

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

---. (2009), "Disagreement as Evidence: The Epistemology of Controversy," *Philosophy Compass* 4: pp. 756-767.

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

---. (2014), "Conciliation, Uniqueness and Rational Toxicity," *Noûs* 50: 584-603.

---. (2016), "Disagreement, Drugs, etc.: from Accuracy to Akrasia," *Episteme* 13: 397-422.

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 University Press).

Decker, J. (2014), "Conciliation and Self-Incrimination," *Erkenntnis* 79: 1099-1134.

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

---. (2010), "How to Disagree about how to Disagree," in Feldman, R. and T. Warfield, eds., *Disagreement* (Oxford University Press).

---. (2013), "The Puzzle of the Unmarked Clock and the New Rational Reflection Principle," *Philosophical Studies* 164: 127-139.

Fantl, J. and M. McGrath (2002), "Evidence, Pragmatics and Justification," *Philosophical Review* 111: 67-94.

Field, H. (2000), "Apriority as an Evaluative Notion," in P. Boghossian and C. Peacocke, eds., *New Essays on the A Priori* (Oxford University Press).

Fleisher, W. (2018), "Rational Endorsement," *Philosophical Studies* 175: 2649–2675.

Frances, B. (2010), "The Reflective Epistemic Renegade," *Philosophy and Phenomenological Research* LXXXI, 2: 419-463.

Gardiner, G. (2018), "Evidentialism and Moral Encroachment," in K. McCain, ed., *Believing in Accordance with the Evidence: New Essays on Evidentialism* (Springer).

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

Goldberg, S. (2013), "Defending Philosophy in the Face of Systematic Disagreement," in D. E. Machuca (ed.) *Disagreement and Skepticism* (Routledge).

Goldman, A. (2001), "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* 63 (1): 85-110.

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

Horowitz, S. (2014), "Epistemic Akrasia," *Noûs* 48:4: 718-744.

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

---. (2010), "Peer Disagreement and Higher-Order Evidence," in Feldman, R. and T. Warfield, eds., *Disagreement* (Oxford University Press).

Lackey, J. (2010), "A Justificationist View of Disagreement's Epistemic Significance," in A. Haddock, A. Millar and D. Pritchard, eds., *Social Epistemology* (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 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: 145-179.

---. (2020), "Enkrasia or Evidentialism? Learning to Love the Mismatch," *Philosophical Studies* 177, 597–632.

Littlejohn, C. (2012), "Disagreement and Defeat," in D. Machuca (ed.) *Disagreement and Skepticism* (Routledge).

---. (2015), "A Note Concerning Conciliationism and Self-Defeat: A Reply to Matheson," *Social Epistemology Review and Reply Collective* 3 (12): 104-112.

---. (2018), "Stop Making Sense? On a Puzzle about Rationality," *Philosophy and Phenomenological Research* 96 (2): 257-275.

---. (2020), "Should we be Dogmatically Conciliatory?" *Philosophical Studies*: 177: 1381–139.

Marušić, B. (2015), *Evidence & Agency* (Oxford: Oxford University Press).

Matheson, J. (2014), "Are Conciliatory Views of Disagreement Self-Defeating?" *Social Epistemology* 29 (2):145-159.

---. (2015), "Epistemic Norms and Self-Defeat: A Reply to Littlejohn," *Social Epistemology Review and Reply Collective* 4 (2): 26-32.

Nelkin, D. (2000), "The Lottery Paradox, Knowledge, and Rationality," *Philosophical Review* 109(3): 373-409.

Neta. R. (2018), "Evidence, Coherence and Epistemic Akrasia," *Episteme* 15, 3: 313 -328.

Nozick, R. (1993), *The Nature of Rationality* (Princeton University Press).

Pittard, J. (2015), "Resolute Conciliationism," *Philosophical Quarterly* 65 (260):442-463.

Reining, S. (2016), "On the Supposed Dilemma of Conciliationism," *Episteme* 13, 3: 305-328.

Rinard, S. (2017), "No Exception for Belief," *Philosophy and Phenomenological Research* 94: 121-143.

Salow, B. (2019), "Elusive Externalism," *Mind* 128: 397-427.

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

Silva, P. (2018). "Explaining enkratic asymmetries: knowledge-first style," *Philosophical Studies* 175: 2907-2930.

Smithies, D. (2012), "Moore's Paradox and the Accessibility of Justification," *Philosophy and Phenomenological Research* 85: 273-300.

---. (2019), *The Epistemic Role of Consciousness* (Oxford University Press).

Sliwa, P and S. Horowitz (2015), "Respecting all the evidence," *Philosophical Studies* 172 (11): 2835-2858.

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

Weatherson, B. (2013), "Disagreements, Philosophical, and Otherwise," in D. Christensen and J. Lackey, eds., *The Epistemology of Disagreement: New Essays* (Oxford University Press).

Weintraub, R. (2013), "Can Steadfast Peer Disagreement be Rational?" *Philosophical Quarterly* 63: 740-759.

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

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