Planning For Failure

by

Robert John Steel

B.A. in Philosophy with honors, Stanford University, 2008

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

PhD in Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Robert Steel

It was defended on

May 16, 2016

and approved by

Michael Caie, PhD, Assistant Professor

Anil Gupta, PhD, Professor

Kieran Setiya, PhD, Assistant Professor

Dissertation Advisor: Karl Schafer, PhD, Professor

Dissertation Advisor: James Shaw, PhD, Assistant Professor

PLANNING FOR FAILURE

Robert Steel, PhD

University of Pittsburgh, 2016

Sometimes I not only judge that P, but I also have some further “higher-order” evidence about my reliability when it comes to this sort of judgment. Perhaps I make an arithmetical judgment, but at the same time I have evidence that I am drunk and so arithmetically unreliable. I argue that the rational response in such a situation is to calibrate one’s confidence in P to the level suggested by one’s higher-order evidence—in this particular case, however confident I would be in P just given the description “I judged that P, drunkenly.” As a special case, this “calibrationist” doctrine entails a conciliatory view of peer disagreement. Some have found conciliatory views to be disturbing, taking them to collapse into psychologism and, at the limit, skepticism. I develop a conciliatory view which need be neither of these things, which is fortunate, since I also argue it is correct.

TABLE OF CONTENTS

PREFACE

This dissertation is dedicated to Samantha Hancox-Li, whose philosophical conversation and personal friendship were invaluable to me during the time spent writing it.

1.0 INTRODUCTION

Later tonight I am going out to dinner. I think: "at some point, I will have to figure out my share of the check. If I drink, then my chances of botching the division go up substantially. Given what I know of my own reliability in these matters, I anticipate two drinks will reduce my chances of getting it right on the first try to a mere 50-50." Fast forward to dinner, where I have two drinks and then I go on to try to divide the check. How confident should I be that I've gotten it right on the first try? I say: I should take it to be 50-50. I should match my current confidence to my earlier estimate.

The view I propound in this dissertation is thus "calibrationist" in character: it requires calibrating one's judgements to one's antecedent estimate of their reliability. This advice furnishes us with a picture of both how to understand "higher-order evidence" and how to respond to it. What is higher-order evidence? Higher-order evidence is evidence on the basis of which one can estimate the reliability of one's judgements. How ought one respond to higher-order evidence? One ought match one's confidences to the estimates formed on its basis.

I hold that following this advice is *all* one needs to do to acquire an *ultima facie* rational belief. An immediate consequence: even if one's initial judgment failed to match the first-order evidence, the belief arrived at by correctly calibrating it ends up rational all the same. As such, on my view there is a sense in which the weight of the first-order evidence turns out to be irrelevant to what one should believe at the end of the day. Namely, we can determine what one should believe just by citing one’s initial judgment and one’s expectation of its reliability, and in so doing we need never say a word about what the first-order evidence actually supported.

Notice, after all, that this is just what I did with the opening check-splitting case. The story I told contains no information about what the correct calculations support; it contains no first-order considerations at all. Yet, nonetheless, I was ready to pronounce on what my ultimate confidence should be. I could do this this because on my view we don’t need to know what the first-order considerations support to know what my ultimate confidence should be.

We can represent this picture somewhat sloganistically with the claim that *rational belief is directly sensitive only to higher-order evidence (and not first-order evidence)*. So sloganized, it’s easy to see the natural alternatives. So, for instance, there’s the mirror image: “right reason” views hold that rational belief is directly sensitive only to first-order evidence (and not higher-order evidence). And then there’s the pluralistic choice: “interactionist” views hold that rational belief is directly sensitive to *both* first- *and* higher-order evidence. Now, this trichotomy is, unfortunately, not exhaustive. However, it does capture much of the current work on higher order evidence and it is the organizing frame on which I hang my dissertation. So:

In chapter 1, I argue against right reason views. I start by raising a fairly standard objection: right reason is extensionally incorrect. The novel contribution, by contrast, comes in my exploration of the following moves and countermoves; I argue that all attempts to finesse the objection fail. I also introduce a theme that will reappear in the subsequent chapters, namely, the thought that there may be multiple epistemic concepts worth caring about. Right reason, I suspect, correctly describes evidential support. But it is a bad view of rational belief. What’s the difference? I take it that the latter concept, unlike the former, builds in right-reason-unfriendly constraints that are grounded in our beliefs about our own efficacy.

In chapter 2, I argue against interactionist views. I argue that they require joint sensitivity to two things—first order and higher order evidence—that could not possibly be 'in play' at the same time. I support this by showing how interactionism leads to akrasia; I diagnose this as stemming from the fact that there is no coherent perspective from which the interactionist advice could reasonable. In putting forward this diagnosis, I also reprise the earlier theme that there are multiple epistemic concepts worth caring about. In keeping with the perspectival metaphor, I allow that there is a perspective from which one can 'see' the first order evidence, and that this perspective is useful for some purposes; and, similarly, there is a perspective from which one can 'see' the higher order evidence, and this is useful for other purposes. What there is not, I maintain, is any perspective from which to see both.

In chapter 3, I move on to giving my central positive argument. I show that calibrationism has the following lovely feature: trying to follow it maximizes your expected epistemic returns, and, as such, calibrationism is the best view *to be committed to.* Now, it may seem suspect, though, to take the desirability of being committed to calibrationism to be a reason to think that it's actually correct; doesn't this confuse the attitude that it's rational *to have* with the attitude that it's rational *to want to have*? I hope that the previous chapters may have primed the reader to be at least somewhat receptive to the answer 'no.' There may be some normative concepts which are constrained by their expected effect when had, and if so this argument shows calibrationists are primed to fill that space. But regardless, even if one thinks there is a deep and crucial conceptual distinction between attitudes it is rational to have and attitudes it is rational to want to have, I do a fair amount of work to show that if our theory routinely has them coming apart there will be hideous consequences with respect to practical rationality, especially when we consider the interlocking effects of multiple decisions made at distinct times. At the very least, we can say that the fact that the calibrationist avoids these hideous consequences counts for something.

2.0 AGAINST RIGHT REASON

2.1 INTRODUCTION

Sometimes we not only have information that directly bears on some important question, but we also have information about the reliability of our own ability to judge that question.

More colorfully, consider *Worried Parent:* my daughter stands accused of a serious crime. I have talked to her about what happened, paid careful attention to her emotional affect as she answered, considered the consistency of her story, and surveyed the physical evidence as it has been gathered by the police. As her parent, I have perhaps more evidence than anyone else when it comes to the question of her potential guilt, for I have a lifetime's worth of relatively close and constant observation to use as a basis for both interpreting and predicting her behavior. But, at the same time, as her parent, I also have a wealth of evidence concerning my (in)ability to reliably and impartially judge her. I know both general features of the ways that parents (mis)judge their children as well as specific aspects of our relationship: facts both like "parents tend to think the best of their children even when they're unwarranted in doing so" and, perhaps, "I have always been fearful and consequently tend to immediately jump to the worst conclusion."

We often have such information; as such it is natural to want to know how it should, or should not, temper our judgments. The practical relevance is especially apparent when we highlight a particular type of evidence about our reliability, evidence due to disagreement, as in *Worrying More:* after spending several agonizing nights surveying my daughter's case, with relief I eventually conclude that she is innocent. However, in a rare honest discussion of the situation with my husband, I discover that he, in his heart of hearts, has come to the opposite conclusion that I have. He believes that she's guilty. I become distressed. I wonder if he is facing the situation more honestly than I am, or if he has a better understanding of our daughter.

One of the things that can call the reliability of our responses to the plain evidence into question is the dissenting judgment of others. But we are surrounded by people who dissent from us on all sorts of issues: not just verdicts of innocence and guilt, but also the economy, good art, how to treat a friend, and whatever else we talk about when we talk about what matters. We are often faced with such dissenting judgments, and they often concern the things in life we care most about, so we should certainly like a normative theory dealing with them. This paper attempts to contribute to such a theory.

I have been putting things in terms of evidence: plain evidence, and evidence of the reliability of our responses to it. Other terms for the same division of evidence include "object-directed" and "reason-directed" (Willenken 2011), as well as "first-order" and "higher-order" (Christensen 2010, Kelly 2005). This evidential framework is natural for pursuing these questions, and so is useful for introducing them, however my discussion needn't actually assume that what is at issue here is exclusively a matter of evidence. Rather, I intend to be maximally ecumenical. So, for the duration of the paper I will instead speak of the interaction between the 'actual justification' and the 'calibration,' and similarly of 'justifying features' and 'calibrating features.'[1] For my purposes here, it's enough to point them out by ostension: actual justification encompasses the features in the worried parent case like my daughter's testimony, affect, the physical evidence, and so on which all bear directly on her guilt; the calibration encompasses all the features in the worried parent case like my emotional bias or the dissent of my husband that bear on my daughter's guilt only indirectly by way of indicating something about my (in)ability to reliably

[1] I take the terminology of 'calibration' and 'calibrationism' from (Schoenfield 2014)

ascertain it.[2]

Given that we often have both some features encompassed in an actual justification and some features encompassed in a calibration, how do those factors interact? In other words, what *should* someone who has both believe, in the epistemic sense of ‘should?’

One bold answer is: those factors do not interact. A person *should* believe whatever they have actual justification to believe. Calibrating features should never either boost or corrode confidence in any such verdict. Call this view the *right reason* view, because it holds that, so long as they’re the product of actually correct reasoning, conclusions should never be further modified in light of calibrating features.

That view has an obvious opposite, and it will be useful to name it for expository purposes. The opposite view claims that the actual justification for a person’s belief is irrelevant to what they should believe; what they should believe is just whatever is supported by the relevant calibrating features. Call this view *calibrationism.*[3]

Although I am sympathetic toward calibrationism, the goal of this paper is not to establish it. Rather, the goal of this paper is to argue against the right reason view. It participates in an argument for calibrationism, though, by eliminating one of the main structural alternatives.

Here’s the plan: in section 2.2, I introduce the right reason view and survey some of its motivations. In section 2.3 I then go on to give my central argument against it: I call this argument the simple argument, and its content is (simply) that the right reason view yields the wrong

[2] See (Christensen 2010) for a fuller characterization, including reasons to think that this type of evidence really is ‘special’ and requires its own treatment.

[3] How do calibrating features, which are not directly *about* the proposition in question, justify any particular level of confidence in it? Roughly: if calibrating features make it rational for me to think I am X likely to be right in my judging P in this type of circumstance, then they supports adopting a credence X in P. Calibrationism is the doctrine that one ought to conform the strength of one’s judgments to their expected reliability. Clearly there is much to be unpacked in this idea; since the topic of this paper is right reason, rather than calibration, I do not enter that task here.

practical results in a wide range of problem cases. Then, in section 2.4, I consider how the right reason theorist might complicate the relationship they posit between the epistemic and the practical so as to block the simple argument; I also lay out some criteria for what would count as a successful instance of that strategy. In sections 2.5 and 2.6 I consider some existing proposals along these lines, but argue that they fail to meet the criteria I've identified. In section 2.7 I outline what I take to be a better, right reason-unfriendly, explanation of the problem cases. Finally, in section 2.8, I consider and respond to a line of objection based on Williamson's anti-luminosity arguments before concluding.

2.2 RIGHT REASON

According to the right reason view, the all-things-considered verdict I should arrive at with respect to my daughter's guilt is only directly sensitive to the actual justification. What do I mean by 'directly sensitive?' I mean: once the actual justification is fixed, that's sufficient to fully determine the verdicts I should arrive at. Any further calibrating features I could go on to introduce—that I am not a good judge, that others disagree, and so on—are irrelevant, at least to the question of what I should believe about my daughter's guilt.

So, that's the right reason view. As before, it's a bold view. Why might someone be attracted to it? It is worth collecting some motivations.

One way to get to the right reasons view is through generally externalist sympathies. So, for instance, one might take one's favored epistemic concept to be best captured by a broad sort of reliabilism: it obtains of beliefs that are produced by methods which are generally reliable. Further, one might think that the most natural understanding of 'methods' to deploy here, the one which preserves best the general motivations for adopting that sort of broad reliabilism in the first place, will be one on which methods do not ‘build in’ higher-order information about their own reliability. That is to say, one in general still counts as using the good, actually reliable method even in the presence of some features indicating that one isn’t.[4] Then one will be led to say that in those cases one’s favored epistemic concept still holds despite the presence of features that indicate that it doesn’t: or, in other words, that actual justification is indefeasible regardless of negative calibrating features.[5]

But one doesn’t need generally externalist sympathies to be drawn towards the right reason view. Suppose one takes one’s favored epistemic concept to be one that is transmitted by good inferences. Take, for instance, any view on which beliefs can be well-founded, and then further beliefs can inherit that well-foundedness by way of being inferred from them by a good method. It seems that everyone, externalistically or internalistically inclined, will count competent deductive inference as a good method, one capable of preserving well-foundedness if any is. But consider a case where one starts with a well-founded belief, and then competently completes a trillion-step logical deduction of one of its consequences. This belief is, by the above description, still well-founded. But also suppose, as is plausible, that no human has ever completed a trillion step deduction without error before, and that there are excellent psychological reasons to take such a thing to be vanishingly unlikely. If so, then although that belief is in fact well-founded, that well-foundedness holds despite the presence of features strongly indicating that it doesn’t: again, the

[4] One may object: even if the presence of negative calibration doesn’t change the method one is using, it may still change that method’s reliability. In general, moves here depend in a relatively fine-grained way on how one is conceiving of methods. As such, though right reason can easily accommodate these thoughts, more would need to be filled in to get the conclusion that it is the *only* way to do so.

[5] (Lasonen-Aarnio 2011) makes this argument. She has also offered similar but broader arguments in her (2014); these drop the appeal to externalism in favor of a much weaker appeal to a rule-driven picture. See footnote no. 34 for further discussion of this argument.

result is that this particular sort of actual justification—justification by competent deduction—is indefeasible regardless of negative calibrating features. And once one takes this sort of actual justification to be indefeasible, then one may be drawn to think the same of actual justification generally.[6]

One can also be driven toward the right reason view by a view on which the extension of one's favored epistemic concept is *a priori*, and that the attitudes someone should have about the *a priori* are not based on calibrating features; perhaps the *a priori* is purely empirically indefeasible, or perhaps it's just that some relevant subset of *a priori* claims is empirically indefeasible. Start with a mathematical example: take someone who, when calculating a 20% tip on a $25 tab at a restaurant, correctly arrives at $5; but then she discovers that her dining partners have all arrived at $6. The line of thought leading to right reason is: that her dining partners have all arrived at $6 isn't evidence against the proposition *that $5 is 20% of $25*. Indeed, no empirical data, about her dining partners or anything else, could be evidence against that, since it's a necessary mathematical truth. But if she hasn't received any evidence against it, then she should be able to continue to believe it. And if she does that, then she can trivially infer that she's right and her dining companions are wrong.[7]

And on some views, what goes there for the mathematical example will go for all inductive

[6] Joshua Schechter (2013) has used long deduction cases like this to argue against closure for precisely the reason that long deductions, even if actually competent, are intuitively risky. The point here is that someone could, by being more attached to single-premise closure, instead be led to taking the modus tollens. Williamson (forthcoming) is an example: he preserves closure by holding that one does know the consequences of one's deductions, but allows that one does not know one knows.

[7] (Weatherson ms.) makes a point like this: Weatherson notes that in (Christensen 2010)'s 'mental math' cases, the calibrationist-style response will require subjects to stop believing things that are literally entailed by their evidence, even though they acquire nothing that looks like direct evidence against the entailment (e.g. they do not acquire evidence that some alternate non-classical logic is correct). He then argues that the only way to make sense of this is to assume, on behalf of the calibrationist, that the entailing evidence is 'screened off' so as to make the entailment no longer available—he objects, though, that this is a bad theory of evidence.

logic more generally. On these views, *all* conditionals relating a total doxastic state to the conclusions it licenses will be *a priori*, and the proposition that some conditional relates some total doxastic state to some conclusion will be similarly indefeasible.[8] So, if one finds these views attractive, then one may be drawn to the right reason view.[9]

What to make of these motivations? I reject the right reason view, but I do not intend to argue against these motivations as such. I think that some of them, particularly the last, are powerful. But it is a mistake to take them to support the right reason view; they are best taken as motivations for thinking there must be *something* in our conceptual toolbox that has a rigid character. My own view is that the concept which best plays this role is the concept of *evidential support*; evidential support is closed under logical entailment and relations of evidential support are both *a priori* and empirically indefeasible.[10] Where the right reason view goes wrong is not in positing such a rigid concept, but in taking its application to line up with the epistemic 'should' of 'should believe;' though we have good reasons to think that such a rigid concept exists and plays an important epistemic role, we also have good reasons to think it is manifestly unsuited to play *that* role. I turn to those reasons now.

[8] The Bayesian family of views are a prominent example, as, for any given set of priors, it is a mathematical truth that conditionalizing them on the basis of some evidence yields a particular result. But also see rationalist solutions to the bootstrapping problem for views on which inductive logic winds up being *a priori*. See particularly (Cohen 2010a).

[9] The idea that propositions about what one ought to believe in a given situation are all *a priori* and indefeasible is central to (Titelbaum 2015)'s argument for right reason.

[10] That evidential support is closed under logical entailment is by no means obvious; for instance, the claim runs into *prima facie* difficulties when considered in light of the paradox of the preface. I will not enter into how to characterize my favored version of closure. Rather, I mention closure not to defend it, but with the more modest goal of illustrate how multiplying our epistemic concepts—for instance, separating *evidential support* from *what one should believe*—allows for some motivations for right reason to be co-opted.

2.3 THE SIMPLE ARGUMENT AGAINST RIGHT REASON

My argument against the right reason view is simple: call it *the simple argument*. The right reason view entails that one's actual justification in some proposition P is indefeasible in the face of any number of adverse calibrating features, and so once actually justified one should continue believing P in the face of any number of adverse calibrating features. I say: it is easy to construct cases in which, due to overwhelmingly many and powerful adverse calibrating features with respect to their belief in P, it's apparent that our subject ought not rely on P in practical reasoning. The best explanation for the fact that they are not justified in relying on P in practical reasoning is that they are not theoretically justified in believing P either; so, calibration undermines actual justification. But that contradicts the right reason view, and so the right reason view is false.

What sort of cases does the simple argument employ? Here's the recipe. Start with a situation in which someone has done whatever is required to actually justify some target belief, so we begin with a *successful reasoner:* [11] Sasha is a civil engineer overseeing the building of a bridge. She has determined, through the standard methods, a schematic for starting construction. The methods she's used are, as noted, standard well-verified ones, and she has succeeded in applying them correctly. Nor does she have any reason to doubt that she has done so. She is justified in ordering construction to go forward if anyone ever is. We may add, if we want, that she knows the bridge is sound.

Then, add a bushel of negative calibrating features, aka *overwhelming adverse calibration:* before she starts construction, her doctor confronts her. He claims to have discovered from recent

[11] These cases can be multiplied endlessly; this one is a close variant of the one found in (Hawthorne & Srinivasan 2013).

tests that she is suffering from a brain lesion that has critically damaged her abstract reasoning abilities; he says that he has consulted with all the relevant experts in his field, double checked the scans, and so on, and has concluded that her ability to do basic calculations is completely compromised (though none of this is, in fact, true, but rather the product of an exceptional mix-up). Upon hearing this disturbing report, Sasha nearly faints; in that moment, she seems to remember having gone sleepless for weeks, a sleeplessness that would have taken a devastating toll on her.[12]

Finally, we add some consequences to make sure the outcome matters, so that the *decision is important:* Through the haze Sasha sees that the foreman is impatiently glaring at her. Time is money. If she orders construction to go forward and it turns out that there are any errors in the bridge schematic, then the bridge will not be structurally sound during the construction process, making a collapse likely. Such a collapse would cost millions of dollars in wasted time and material and would be likely to kill several workers. The foreman wants to know—start building or no?

What should Sasha do? I think it's difficult to deny that there is some sense in which the answer is: she should not order construction to go forward. To do so would be a terrible, irresponsible choice. The thought of someone making that choice, in the face of learning all that Sasha has learned, is disturbing.

But what sense is this, the sense in which she ought not order construction forward? It clearly cannot be the 'objective' sense of ought. The objective ought tracks the choice which would, in fact, produce the best outcome. Since (at at least one point) Sasha was able to know that the bridge would be sound, the factivity of knowledge implies the bridge would be sound. On

[12] On some views, she could not actually have gone sleepless without thereby necessarily degrading her reliability and hence rendering her initial actual justification impossible. But on all views she could receive strong, though misleading, evidence that she had gone sleepless even if she had not.

natural background assumptions, it follows that ordering its construction forward is the course of action that would, in actuality, yield the best results. So in the objective sense it's just trivial that Sasha ought to order construction forward.

So, when we are trying to divine the sense of 'ought' in which she ought not order the bridge forward, it is natural to look instead to the 'subjective' ought. This sense (or, perhaps, family of senses) does not hinge on which of the courses of action available would *actually* have the best results if Sasha were to choose it, but rather, on which course of action looks best 'from her point of view,' however we are to understand that in a specific instance. These come apart as, for instance, when putting all her money on black would in fact win her the current spin of roulette, given the current velocity, position, and etc. of the ball and table—but, of course, she doesn't have the relevant sort of access to that information. So although putting all her money on black is objectively best, it is not subjectively best; in the case of roulette, barring exceptional circumstances, what is objectively best (betting on the winning color) is never subjectively best (because, given the odds, it is subjectively best not to bet at all).

As the case of roulette suggests, subjective 'oughts' align more neatly with a cluster of notions surrounding rationality: intelligibility, planning, and praise and blame, and so on. When a person abstains from betting in roulette, their prudential success is more rationally intelligible than the success of the person who spontaneously bets on the winning color; we take that spontaneous winner who bets every round to be criticizable for their imprudence, even when it leads to an actual good outcome; when we consider betting ourselves, we plan to abstain; and so on. I don't want to rest too much on any particular one of these connections, because I don't think it's in the end theory-neutral either which wind up being vindicated or in what way. Rather, what's important is that the cluster helps characterize the type of concern we are taking up.

So, the sense in which Sasha ought not proceed with construction is a subjective one. Can we characterize it any further? I think the answer is yes. In my view, the sense in which Sasha ought not order the bridge forward is this: if she believed what she should believe, then relative to those beliefs ordering construction on the bridge forward would look poor. Because she ought not believe the bridge is sound, she ought not order construction forward. Since this contradicts the right reason view, its proponents must either deny that there is *any* sense in which Sasha shouldn't order construction forward, or they must find an alternate understanding of the 'should' at issue such that it is compatible with right reason.

We may thus summarize the simple argument as follows: when we take a reasoner who has successfully reasoned to some conclusion, add overwhelming adverse calibration, and then ask "what should they *do*, in a situation in which the truth of that conclusion is very important?" we get the answer: they should not rely on that conclusion. Furthermore, the best explanation for why they ought not rely on that conclusion in the practical context is that they ought not believe it. But the right reason view says they *should* believe it. So, the right reason view is false.

As the name implies, the simple argument itself is not particularly complicated. Nor is it novel; others have suggested in passing that the beliefs the right reason view recommends seem especially dubious when we consider someone actually acting on them (e.g., Christensen 2010, Schechter 2013). The point of the presentation here is to take the argument implicit in that observation and make it explicit as possible.

The argument may be simple, but evaluating its success is not. That is the task to which I turn now. I will begin by addressing existing responses that accept the basic intuition about what Sasha ought to do—namely, not order construction to go forward—but which then try to block the explanatory step to the conclusion that she ought not believe the bridge sound either. They do so by introducing some mix of new terms, distinctions, and principles which complicate the relationship between what Sasha should believe and what she should do, with the goal being to find a way to let the claim that Sasha should believe the bridge is safe coexist in harmony with the intuition that she would definitely be wrong to straightforwardly act on that belief.

2.4 BLOCKING THE SIMPLE ARGUMENT: CRITERIA OF SUCCESS

Ordinarily, if we ought to believe *p* then we ought to act as if *p*. And so, just by contraposition, it's also true that if we're pretty sure that we oughtn't act as if *p* then we can also be pretty sure that we shouldn't believe *p*. These conditionals, just taken materially, represent a default 'matching' between the practical and epistemic: it's not that they cannot fail, but just that if they do we should expect an explanation for why. So, we can examine the question of whether they hold in Sasha's case by way of examining whether there's any explanation for why they wouldn't: are there special features of Sasha's case, and others like it, which complicate the straightforward relation?

It will be useful to introduce this idea by way of some remarks from Williamson, in response to a similar problem his system faces (Williamson 2005: 479-483). In that system, your evidence is equated with everything you know, and so the evidential probability of any particular thing you know—the probability of it conditional on your evidence—is always 1. Furthermore, Williamson wants to claim that we know some things. But on standard decision theory, believing propositions with probability 1 leads to what intuitively seem like irresponsibly incautious choices. So there is a version of the simple argument here: these choices seem wrong, so the epistemic theory which generates them is wrong too.

In trying to defuse this worry, Williamson begins by noting that a version of this problem already exists entirely independently of his system's particular feature of treating knowledge as evidence. That's because standard decision theory independently requires assigning probability 1 to all logical truths, and this is already enough to generate the intuitively objectionable choice behavior. He claims that the solution, in both cases, is to distinguish the theory of rationality from what he terms 'good cognitive habits.' The theory of rationality may recommend assigning probability 1 to all logical truths and then being ready to wager any amount, however outrageous, on one's correctness, however, reasonable humans with good cognitive habits will not try to follow that advice. Rather, they will acknowledge that they may have made a computational error and will take steps to manage their fallibility.

As such, someone with good cognitive habits may do something that is in fact irrational. They may fail to take a bet, even though that bet is guaranteed to succeed because it is on what is in fact a logical truth—if the costs of failure are high enough, and they are prudently managing their own fallibility. And they may do a similar thing when it comes to what they know. They may, irrationally, fail to take a bet, even though they know that the condition they are betting on obtains. Furthermore, they may not only do something which is in fact irrational, but they may also do something they *know* to be irrational, or something they know they know to be irrational, and so on.

That what is rational can come apart from what the person with good cognitive habits chooses does not implicate either the theory of rationality or the cognitive habits in question as wrong. They are just different subjects, and there are reasons internal to the theory of rationality to make it demanding in the way that generates this disconnect. We confuse the two subjects when we take our intuitions about objectionable betting behavior to tell against any particular theory of rationality; what those intuitions track are instead the good cognitive habits of reasonable people. Or, so his argument goes.

Here we are interested in the right reason view specifically; does this argument suffice to defend it?[13] Williamson's discussion has the right potential shape: it makes a distinction that purports to show how claims about objectionable choice behavior can be dissociated from claims about rational belief; since the simple argument relies on an explanation of the badness of choices in terms of the badness of beliefs, this could serve to undermine it. But despite this promising shape, closer examination reveals significantly more work must be done before this could be turned into a successful response.

Consider: we have three prima facie normative notions on the table here, 'rationality,' 'good cognitive habits,' and 'the epistemic 'should' of 'should believe.'' The right reason view is defined in terms of the last. It doesn't just claim that there is *some* rigid epistemic concept in our toolbox; as I've already indicated, I myself think that much is true. Rather, the right reason view's distinctive claim is *that the 'should' of 'should believe'* is such a rigid concept, and so Sasha *should believe* according to her actual justification without respect to calibrating features. In order for the distinction between 'rationality' and 'good cognitive habits' to do any work in defending that view, we need an explanation of how those notions, and their normative import, relate to that of the 'should' under investigation.

The desired result, which would put the right reason view in the clear, would be that 'should believe' winds up lining up with 'rationality,' whereas our intuitions about what ought to be done

[13] Williamson endorses the right reason view for at least one class of cases: closure cases, as noted in footnote no. 6. There he holds that one knows, though one fails to know one knows, and though one may even correctly take the objective chance of what one knows to be very low. These cases can easily be described in terms of the recipe of the simple argument. So although Williamson offered this discussion, as noted, in a slightly different context, he seems committed to its general applicability.

wind up lining up with ‘good cognitive habits.’ Then, proponents could maintain that they were right all along about their thesis concerning what Sasha should believe, and to the extent that the seemingly-contrary intuitions about choice embodied in the simple argument are about anything it turns out to be an entirely different topic.

But in order to know whether these concepts line up in the required way, we actually need an account of *what* good cognitive habits require, *in what sense* they require them, and *why* they require them; it cannot simply be left open, ready to be appealed to as necessary to fix up problem cases.

Why do we need such a spelled out account—why does a promissory note, or general gesture in the direction not suffice? Because until one is provided, we’ll be left with a lurking worry. Suppose that what someone with good cognitive habits does directly explains, in *every* case, what we think a person ought to do, and what a person knows, or their actual justification, only ever influences what a person ought to do by way of mediately affecting what the person with good cognitive habits would do. If these conditions hold, then it will look like good cognitive habits were the interesting notion all along, the one that we were trying to get at when we asked questions about what the ‘should’ of ‘should believe’ attaches to in these situations. And so positing good cognitive habits would not be much of a defense of the right reason view; it would defend it only by ceding the topic.[14]

The upshot here is that Williamson’s discussion points us in the direction of a solution for

[14] Williamson considers the person who, upon reflecting on the difference between rationality and good cognitive habits, decides that good cognitive habits are more interesting. His main response is that such habits will be ‘non-luminous’ just as much as rationality is (Williamson 2005: 483). It’s not clear how this helps, though. Williamson is committed to the explanatory centrality of knowledge to a whole host of phenomena. Good cognitive habits could threaten to usurp that centrality without being luminous—after all, it is central to Williamson’s argument that these roles do not require luminosity. See Williamson (2000) for his seminal exposition of his knowledge-first view. I further discuss luminosity, and why it seems to me beside the point, in Section 2.8

the right reason theorist, but also to a constraint on what such a solution must look like in order to be successful. The right reason theorist can undermine the simple argument by introducing additional concepts which complicate the interface between belief and choice; but at the same time, for whatever concept is introduced it must be clear that it does not threaten to supplant rational belief as the real match to the epistemic 'should' we initially took as our topic—for if it did, then whatever norms governed that concept, be they calibrationist or anything else, those would look to be the norms we were trying to investigate all along.

That is one constraint; here is another obvious one. If the introduction of some new concept is to undermine the simple argument, it must block all of its putative instances. So there must be no cases of the recipe left over to stand as counterexamples.

I now turn to some proposals in the literature; we can think of these as ways of trying to fill in what 'good cognitive habits' amount to in such a way as to secure the desired results. I will argue that although these attempts can variously meet either of the two constraints I've outlined, they cannot meet both at once. To the extent that they capture all the cases, they do so by introducing another concept that seems to be a better match for the epistemic 'should' we are interested in.

2.5 BLOCKING THE SIMPLE ARGUMENT: INCONGRUOUS HIGHER-ORDER BELIEFS; PRACTICAL BRIDGE PRINCIPLES

I begin with Weatherson's proposal for handling the simple argument (Weatherson ms.). It has two components. The first is epistemic: it appeals to higher-order belief. The second is practical: it appeals to special features of Sasha's case, like the fact that she may have responsibilities to others. Each of these components is a natural resource for the defender of right reason, so it is worthwhile to see what can be done with them. I will argue that Weatherson's particular use of them fails as a defense against the simple argument: roughly, because they cannot explain practically simple cases. But what I take to emerge from the discussion is a stronger verdict than just that—not only does Weatherson's particular use of these resources to defend against the simple argument fail, but it fails precisely because they are generally inadequate to the task.

Before we get to that, though, we must get the proposal on the table. Begin with the epistemic side of the story. Weatherson treats cases like Sasha's—or, at least, some of them[15]—in the following way: Sasha's actual justification continues, throughout, to support the conclusion that the bridge is sound, and so she should continue to believe that first-order claim. What changes when she acquires the overwhelmingly adverse calibration is that she gets evidence against the higher-order claim 'I should believe the bridge is sound.' The disagreement of experts, the fatigue-induced confusion, and so on, all contribute evidence that she is not believing as she should. So she should respond to that evidence by doing as it suggests, namely no longer believing that she should believe the bridge is sound, or perhaps even believing that she shouldn't. But the rebutting of that higher-order claim leaves untouched her first-order justification. So, in the end, she should believe both 'the bridge is sound' (as her actual justification supports) and 'I shouldn't believe the bridge is sound' (as her adverse calibration supports). Each of these beliefs is supported by the evidence that bears on it, so in holding both she responds properly to the evidence she has.

[15] His discussion in his section 2.1 centers on basic inferences; he concludes it by saying that 'at least some of the time' cases may meet the description he outlines, and a natural reading of that 'sometimes' is as a restriction to the basic inferences. But then in his discussion of the Sasha-like 'sleepy hospital' case in his section 2.2, which I am now considering, he neither stipulates nor argues that the doctor's inference is basic. I think my criticisms will equally apply regardless of whether we take the inferences in question to be basic, so I do not labor this interpretive point. (Weatherson ms.).

This combination may look to be incoherent: Sasha should believe the bridge is sound, even though she should believe she shouldn't? Isn't this sort of akrasia objectionable?[16] But Weatherson defends against this charge, and it is not the interesting point of contention here. Rather, simply grant the possibility of such 'rational mismatches' between first-order and higher-order belief; to introduce some terminology, call the mismatched higher-order beliefs 'incongruous.' My interest is in whether an appeal to incongruous higher-order beliefs could defuse the simple argument.

On this score, it is worth noting that positing the existence of incongruous higher-order beliefs does little just on its own; we also need an explanation of their practical significance. Recall that the simple argument is, at heart, an explanatory argument: right reason cannot explain why Sasha ought not order construction forward. In order to resist this argument, right reason must not just have *some* explanation, but it must have an adequate one. Thus, what we are looking for is a plausible account of how higher-order beliefs become practically salient, such that, when they are incongruous, they can participate in a good competing explanation for what Sasha ought to do.

Fortunately, Weatherson goes on to spell out just such an account. This is where we get the distinctly practical part of the story. It goes as follows: as described above Sasha is a civil engineer and people's lives depend on her competent execution of her job, so she has certain institutional responsibilities which may limit the courses of action she can reasonably choose. It's also true that, in the case as described, there is 'safe' and a 'dangerous' choice. Sasha could always wait and double check; even if doing so would have costs, those costs would be guaranteed to be much less than those of a collapse. Weatherson holds that general principles of caution and special

[16] For a compelling case that epistemic akrasia is indeed, with rare exception, irrational, see (Horowitz 2014).

institutional obligations complicate the interface between theoretical and practical rationality in Sasha's case, and that they do so by imposing additional requirements on her. Sasha ought not order construction on the bridge forward because although she knows it is safe, she fails to know that she knows—and, since ordering the bridge forward is incautious, and may violate an institutional responsibility she has for others' safety, the epistemic 'bar' for action is set higher than it would otherwise be; it is set high enough to require the additional higher-order knowledge which she lacks.

Refer to any principle that first picks out some special practical features of a choice situation, and then correspondingly complicate the way one's epistemic profile participates in one's decision, as a 'practical-epistemic bridge principle.' Can the practical-epistemic bridge principle that Weatherson proposes here—the principle that institutional roles plus norms of caution require higher-order justification for risky action—defend the right reason view from the simple argument? How does this response do *vis a vis* the criteria I outlined? It easily passes the first. The principle under consideration has clearly defined content, and from that content it is apparent that there is no worry it will wind up simply redescribing the real theory of what one should believe. An institutional obligation, for instance, to treat all defendants as innocent until proven guilty just has no relation to—and certainly does not require—the claim that one *should think,* during a trial, that all defendants are innocent.

But though it passes the first criterion easily, it seems to me that it correspondingly cannot pass the second. For it to pass the second, it would need to be that there's some explanation along these lines for not just some, but all of the cases where the combination of *successful reasoning, adverse calibration,* and *an important decision* make the right reasons view look like it suggests some calamitous decision. But it's easy to use the basic recipe to generate cases that involve neither institutional obligations nor general issues of caution. When we do, we find that when those cases are constructed by the same formula they yield the same result.

The particular case I will use takes place against some background statistical information; this is not an essential feature, but it helps to cleanly push apart the verdicts of the right reason view from calibration-sensitive alternatives. Given that I appeal to such information, though, it is helpful to start with a preliminary observation about how it works. Suppose that someone tells you: I have a fair 20-sided die. It has a 20 on one of its faces, and a 1 on every other. If they tell you that, then you can reason that on any given roll there is a 5% chance of rolling a 20 and a 95% chance of rolling a 1. Nonetheless, if you *see* them roll it, and you see a 20 come up, you can become very confident that it has just rolled a 20: much more than 5%, certainly. General statistical information about how the die rolls, within normal bounds, is screened off once one can see the result in the actual case at hand.[17]

Now to the case: imagine that we have a machine that, on command, spits out a well-formed formula in propositional logic. I know as statistical background information about this machine that it produces a valid schema in only 5% of cases, with the remaining 95% being invalid; if we like, we can imagine that I know it uses a fair 20-sided die to decide which to do.

This odd machine is the centerpiece of a popular game show, TAUT OR NOT, on which I am currently appearing as the contestant. My job is to stand on a podium, press the button on the machine, and be confronted by the formula it produces. I then say yea or nea on whether that formula is a tautological schema. If I answer correctly, I get $10,000; otherwise, $0.

[17] This fails to hold in cases where the statistical information is not within normal bounds, as in e.g. Williamson’s (2000: 205) example of a statistical miracle during a long series of draws with replacement; nothing in this example turns on the proper treatment of the extreme case. Insofar as I can see, if the extreme case is different it only serves to strengthen the argument against the right reason view.

I am confronted with the following schema: $P \rightarrow (Q \rightarrow ((P \vee Q) \rightarrow {\sim}({\sim}((P \vee P) \rightarrow P) \;\&\; {\sim}P \;\&\; {\sim}Q)))$. I then scrawl up some truth tables and, on that basis, I correctly deduce that it is true on all assignments. However, at the same time, I know that I have performed very poorly on similar tasks in the past. In my baby-logic class, which is the furthest I've ever gotten, I've routinely messed up my truth tables. Furthermore, at the moment I happen to be very tired, having been on an all-night bender. In fact, noticing my sweaty palms, I reasonably come to believe that the drugs aren't even out of my system yet. I feel faint, etc. etc. Pile on the negative calibrating features as appropriate—avail yourself of whichever disturb you most. Keep going until it seems positively *insane* for me to remain confident in my reasoning.

In the case as described, we have an actual justification consisting in a competent deduction. On all plausible views, having competently deduced a conclusion is a good actual justification. So, in order to make this into a bad case, as per our recipe, we then took that good justification and added in a heap of negative calibration. Now ask: what should I do?

I say: I should indeed become extremely unconfident in my reasoning; I should think it very likely that I've messed up my tables somewhere, and thus that they count for little. Once I'm not counting my tables for anything, what I have left is just my knowledge of the general statistical behavior of the device. Since I know it produces tautologies only 5% of the time, and that's all I have to go on, I thereby ought to answer that the formula is not a tautology. In this case, as in bad cases generally, I explain why I ought not answer the formula is a tautology by citing the fact that I ought not believe the formula is a tautology.

By contrast: what does the right reason view say? It cannot say that I ought not believe the formula is a tautology. On the right reason view, actual justification cannot be defeated by any amount of calibrating features; since correct deduction by way of a truth table provides actual justification, it follows that the right reason view is committed to holding that said actual justification remains undefeated. And as long as that actual justification is present, then background statistical information about the performance of the device should be screened off: as far as the right reason view is concerned, our situation is like that of person that knows the die rarely rolls 20, but then sees it roll 20. So in terms of what I should believe, the right reason view gives the verdict that I ought to believe it's a tautology.

Can it nonetheless recover the answer that I ought not answer the formula is a tautology—perhaps by citing the demands of caution, or institutional responsibility? I say: no. The case has been constructed so as to render these factors irrelevant. The fact that I ought not answer the formula is a tautology is not explicable by practical obligations, because it is a case of pure prudence—it's only my own future money at risk. Nor can it be explained in terms of a principle of caution, since the payoffs to being right or wrong for each possible answer are symmetrical. So here we have what I take to be just yet another case following the basic recipe, but one that cannot be explained by Weatherson's practical-epistemic bridge principle. So, we get the conclusion that considerations of caution and responsibility cannot block all the cases: this one stands over as a counterexample.

As a conclusion, this may seem rather unexciting: even if the specific practical-epistemic bridge principle Weatherson outlined cannot get a grip here, surely there are many such potential principles. Perhaps we just need to find the right one. But I think that's the wrong lesson. Not only is this a case where norms of caution and institutional responsibility don't apply, it's a case where it's hard to see how *any* complicating practical-epistemic norm could enter the picture. It is a case of pure prudence, one which is all upside and no risk; in such cases the only relevant norm is maximization as informed by one's confidences. So there's just nothing complicated about the interface between the practical and epistemic here, and so there are no features for complicating practical-epistemic principles to latch onto. Given that even cases like this can be bad, there are going to be bad cases that will go unexplained whatever practical-epistemic principles we might favor.

What's more, I take this to also be bad news for any appeal to incongruous higher-order beliefs. Recall that we started by allowing the possibility of rational mismatches, and bridge principles came in with the purpose of describing how the attendant incongruous higher-order beliefs were supposed to rise to practical salience; we needed an explanation of why and how higher-order belief should be practically relevant in the just the ways necessary to defend right reason. What we have found is not just that bridge principles can't do that work, but that the reason that they can't is because right reason faces counterexamples even in cases where the interface between practical and theoretical rationality is *maximally simple.* But once we see that practical defeat persists even in such maximally simple cases, it casts doubt on the general explanatory power of incongruous higher-order beliefs.

Of course, I can't demonstrate the inadequacy of appeals to higher-order beliefs by enumerating and rejecting all conceivable such explanations. But nonetheless, I can say something that, in my view, will very generally count against them. Namely: it is hard to resist the thought that the most basic way our beliefs rationally contribute to our actions is by representing the world to be particular ways, such that we then choose among potential actions on the basis of how good doing them would be, should the world be as our beliefs represent.[18] But notice: when I am on the logic game show, trying to decide whether to answer that the formula is a tautology, I am utterly

[18] I intend this formulation to be independent of any particular detailed development, as in, e.g., causal or evidential decision theory; rather, I take it to be a platitude common to both.

indifferent whether the world is such that *it's a tautology* or whether it is additionally such that *it's a tautology and I ought not believe that it is*; I get the exact same payout either way. But then, adding the incongruous higher-order belief to my belief set doesn't actually represent the world as being any more hostile to my choice; indeed, after adding the incongruous higher-order belief, my total beliefs still represent the world as decisively favorable to answering it is a tautology and decisively hostile to answering it isn't. Then there is a mystery in how adding the incongruous higher-order belief could rationally explain why I ought to change my choice.

I have been fairly generous in allowing that there may be special practical circumstances in which the basic relation between belief and action outlined above fails to hold; perhaps, when complicating considerations regarding caution and institutional responsibilities are brought into the mix, these factors can independently explain why adding that incongruous belief rationally requires me to change my choice. However, we have found that the simple argument can proceed even from cases not involving any special complications. So we need a different explanation, for those cases, of why incongruous higher-order beliefs should be practically salient. And reflection on the content of these beliefs gives us reason to be skeptical that any right-reason friendly explanation is possible.

Of course, there is one very obvious explanation of why incongruous higher-order beliefs should be salient to my choice: namely, because acquiring the belief 'I ought not believe it's a tautology' rationally requires me to abandon my belief that it's a tautology, and *that* change in first-order belief is clearly relevant to my choice. But this is just the answer the right reason theorist is barred from giving, for they deny the change in first-order belief. Rather, it is the explanation posited by the simple argument: it is the simple argument that explains the effect of overwhelming adverse calibration on choice in terms that ultimately go through first-order belief. What right reason needs is a *different* explanation, one that is equally as good. And what I take the above considerations to highlight is how bad the explanation in terms of incongruous higher-order beliefs looks in comparison, when taken on its own. The claim that incongruous higher-order beliefs are *directly* relevant to choice *not* by way of influencing first order beliefs is just mysterious: after all, once we hold fixed the first-order beliefs the additional representational content of the incongruous higher-order beliefs is irrelevant.

I conclude, then, that neither the appeal to incongruous higher-order beliefs nor the appeal to practical-epistemic bridge principles can serve to defuse the simple argument. We see this by focusing on cases where the interface between practical and epistemic rationality is at its simplest. In these cases, it's still true that we ought to respond to overwhelming adverse calibration by modifying our choice behavior. But yet, complicating practical-epistemic bridge principles get no grip; and given that there is no complicating story, there is correspondingly no explanation for why incongruous higher-order beliefs should be relevant—holding that *they just are* would not be a form of explanatory progress. So, we remain lacking an alternate explanation that could plausibly compete with the one put forward the simple argument.

2.6 BLOCKING THE SIMPLE ARGUMENT: REASONABLE SUBJECTS

Independently plausible practical-epistemic bridge principles—relating to the fact that Sasha is an engineer, or to the fact that there is a clear cautious and incautious choice—seem unlikely to get far in defending the right reason view. As above, they just aren't going to get all the problem cases; that those features are present in many of the examples from the literature is an accidental rather than essential feature of the underlying phenomena.[19] And given that such principles fail, it looks like appeals to incongruous higher-order belief are also in trouble; given what they are about, it's hard to explain why they'd matter. So what we're looking for is something that's present in all cases, but yet at the same time looks like it's always poised to matter.

A proposal: perhaps something significant that's common across *all* the problem cases is that they involve a kind of bad habit. Someone who acts as the right reason view apparently recommends, although they will be successful in that individual case, manifests a kind of disposition that will get them into trouble elsewhere. So, for instance, Lasonen-Aarnio introduces the notion of a *reasonable subject:* a reasonable subject is one who adopts policies and maintains dispositions that are overall conducive to acquiring and maintaining knowledge, and a reasonable belief is one which is a manifestation of these dispositions to know (Lasonen-Aarnio 2011).[20] Subjects that act in unreasonable ways are epistemically criticizable. They have adopted methods that are not good methods to adopt, methods that come with bad dispositions. However, though clearly related to knowledge, reasonability is not a condition on knowledge. Knowledge, like everything else, can sometimes be acquired by strategies that are not generally good strategies for acquiring it.

To return to Sasha's case, suppose Sasha sticks with her initial evaluation even in the face of the negative calibration. On this view, if she did so she could retain her knowledge. However,

[19] One may wonder: if it is not triggering something like a principle of caution, what is the point of the high stakes in the bridge case? What has been the point of consistently filling in all our scenarios with substantial practical consequences? I do so in order to heighten and sharpen intuitions. In low-stakes cases, especially those that are purely self-regarding, I think our intuitions are relatively weak and may be swamped by side questions (e.g., I suspect we may take subjects to have a *moral right* to their own judgment in low-stakes self-regarding cases, and that this may be running interference on our readiness to issue criticism).

[20] Hawthorne and Stanley (2008) suggest a similar distinction between evaluation in terms of reasoning as one ought and features of one's 'epistemic character' as a strategy for defending their knowledge-based norm for action; and both of these are ways of filling in Williamson's general "good cognitive habits" more concretely. I find all these proposals dissatisfying in just the same ways to be discussed.

she would be acting unreasonably. Sticking to her guns would manifest a disposition that, though it preserves knowledge in this case, would generate bad results in many others—others where she merely *thought* she was preserving her knowledge, but was in fact dogmatically clinging to a false belief. Our urge to say that Sasha doesn't know is a misplaced product of our recognition of her unreasonability.

What's to be said about this line of response? That all depends on the normative work that unreasonability and knowledge are supposed to be doing, respectively. It is worth emphasizing at the outset that there is a possible division of labor here which is knowledge-friendly, structurally quite close to my own view, and not disqualified by the arguments of this paper.

On that version, unreasonability is the normative notion which most closely lines up with the epistemic 'should' which was our topic. What a person should or shouldn't believe turns on what the reasonable person in their situation would believe, not directly on what they are in a position to know; sometimes, as in Sasha's case, they may be in a position to know something even though they shouldn't believe it, because it is not what the reasonable person would believe. What emerges is a sort of virtue epistemology with the reasonable person playing the role of the *phronimos*. On such a picture, knowledge may still play an important explanatory and metaphysical role, by virtue of giving content to the idea of the reasonable person—the reasonable person's dispositions are those that effectively aim *at knowledge*. But it would still not itself be the normative notion which directly explains what one should believe.

This picture, whatever its advantages, is *not* a version of the right reason view. It aligns the epistemic 'should' under examination with what the reasonable person believes, and what the reasonable person believes is not wholly determined by their actual justification—rather, it is sensitive to calibrating features. Thereby, this view exhibits the structure I insist is required and which the right reason view is defined in opposition to.[21]

So there is a certain kind of normative work that reasonability cannot do for the defender of the right reason view. It cannot be the direct normative standard for what one should believe, where by that I mean that it cannot generally make valid instances of the schema 'it is reasonable in C to believe P, therefore so-and-so in C should believe P.' Let us hold fixed that we not expect reasonability to do that work, and then ask how else it may help. I turn now to the suggestion that the work it does may be just contributing a certain sort of value.

For the sake of argument, we can allow being reasonable any combination of types of value, i.e. practical, epistemic, instrumental, and final. And once we do, we may have new resources that, as before, could potentially help us complicate the relation between the epistemic and the practical. So, we might say: if Sasha believed the bridge was sound, she would instantiate a certain form of disvalue. This disvalue may be enough to make it such that, on the balance, she shouldn't order construction on the bridge to go forward. And this is true even if she still *should believe* the bridge is sound. So perhaps unreasonability can thereby do the work we tried to do earlier with principles of caution, institutional responsibilities and the like—perhaps it can secure the desired practical judgment on some basis that doesn't implicate the falsity of right reason.

This line strikes me as fairly hopeless, however. Sometimes there is nothing wrong with accepting some disvalue, when it promotes another greater value. As such, adding in reasonability as a thing of value does not secure the desired result, namely that Sasha ought not order construction on the bridge forward.

To illustrate: suppose I am a novice rock climber. Since I am a novice, maintaining good

[21] Given both her emphasis on unreasonable methods being bad for us to adopt in section 5 of her (2011), and genuinely criticizable in sections 5 and 7 of her (2011) and also at (2014: 343), I suspect that Lasonen-Aarnio may ultimately have a picture like this. But then again, she also gestures toward an error theory; see footnote no. 22.

form takes much painstaking effort and is very slow going. If I climb my fastest, it is only by using poor form; and, like most such skills, using poor form reinforces poor form. Poor form has some costs: it makes my future climbs less rewarding and more dangerous, and we may even add that it is simply bad in itself, a lack of human excellence. But suppose on a particular occasion I happen to see a distressed person, clearly injured, lying in a crevasse. I know that if I climb toward them at my fastest pace I will thereby use and reinforce poor form. But in this case the importance of reaching this person quickly, to provide vital aid, clearly dwarfs any considerations of maintaining my rock-climbing skill. If, out of a concern for preserving my form, I instead took a slow and painstaking route to get to them I would exhibit a perverse misvaluation of the relevant features of the situation. My form just doesn't matter that much. Even if it somehow meant I could never climb again, the right thing for me to do would be to prioritize rescuing the person in distress.

If what were at stake in Sasha's case really were the goodness of her intellectual form, then similar remarks would apply. The value of having good intellectual form is just less important than many of the goods one might be reasoning about. Delaying construction, for instance, leads to waste and cost overruns; for a sufficiently large project, like many bridges, those costs will be far more significant than the degradation of some individual's intellectual form. We can make this especially acute by supposing that Sasha is retiring, and that this is her last job. Her intellectual habits will not be relevant to any future costly, potentially dangerous construction. If that is so, then she should be able to reason as follows: since the case has all sorts of negative calibrating features, it follows that by ordering construction forward in the face of them I will thereby acquire and reinforce a bad intellectual habit. But my judgement is in fact right in this instance, and the bridge being built on time and under budget matters more than my intellectual habits being any good, so I ought to order construction forward anyway.

If reasonability were merely another things of value, then her reasoning here should be unimpeachable. After all, similar reasoning is unimpeachable in the rock climbing case. In cases where a lot is on the line, cultivating intellectual virtue may matter much less than getting immediate results.

But I take it that this doesn't track our judgments with respect to the case: even when she is about to retire, Sasha's decision to order construction on the bridge to go forward would be wrong. The reason it's wrong is because, in the face of such overwhelming evidence of her error, she should be worried that *in this very case* the bridge is unsafe. Or, at least, I submit that's what we're actually judging, and if I am right then this is an effect that cannot be captured by adding reasonability as just another value in the mix. Just adding reasonability as another value leads to the result that as the other potential consequences grow in significance, then the relative significance of reasonability should wane. But this is to say that when the stakes are highest and it matters most, you should pay the least attention to negative calibrating features. This, however, seems like the wrong prediction: rather, it is in the high stakes cases where we have the clearest judgment that it would be wrong for Sasha to proceed even in the face of such serious evidence of her own error. Adding reasonability as a value not only cannot get the central result we wanted, but it makes bad predictions about the relevant factors that are generating that result.[22]

[22] A last suggestion: perhaps an appeal to reasonability cannot allow right reason to secure the relevant practical judgments, but it might allow right reason to nonetheless account for them. The most radical option for the advocate of right reason is to simply deny that these judgments are correct. That is to say, they might hold that Sasha ought to order the bridge forward after all. On its own, this is quite difficult to swallow. But one might think that invoking reasonability—which, we may still allow, has some value in many ordinary contexts, which often goes hand in hand with what one should believe—now gives them an explanation of why we mistakenly tend to think otherwise: it may be that we do so because we have accidentally focused on what is reasonable to think and do, and confused that with what she ought to think and do. (Lasonen-Aarnio 2011: 6) explains why we may be especially prone to confuse reasonability with knowledge; perhaps this story can be given mutatis mutandis for what is reasonable to believe and what we ought to believe.

The upshot, then, is this: if being reasonable is given a central normative role, such that what one should believe is what the reasonable person would believe, then the resulting view no longer counts as a version of the right reason view. But if, on the other hand, being reasonable is treated as a merely another thing of value, then it is inadequate to vindicate the relevant practical judgments. Since they are apparently robust against being explained away, we really ought be in the market for an account that vindicates them.

2.7 REGULATORY STATES

I take it that the simple argument is *prima facie* compelling. If I am right that extant defenses against it fail, then we have reason to believe the simple argument's *prima facie* weight survives *ultima facie*. I take myself, then, to have already offered an adequate argument for my core thesis: namely, that the right reason view is false.

Nonetheless, it is unsatisfying to leave it at that. The simple argument refutes by counterexample. Such arguments can be very good for showing us *that* a view is wrong, but often cast little light on *why* it is wrong, or *how* it went wrong. In the present case, these latter questions still call out for answers, and so working toward answers for them will be the aim of this section. My strategy is to introduce a competing positive view—my own preferred understanding of the epistemic 'should.' I hold that this view can co-opt some of the motivations of right reason, while

Ideally, an error theory pointing out an alleged confusion like this would be such that, upon being apprised of it, one's urge to make the diagnosed error diminished. I, at least, find the impetus to judge that Sasha ought not order construction forward to be as strong as ever. As such, I think we should only agree that this is an error after all if there really is no alternative. I consider and argue against what I take to be the most prominent reason to think there is no alternative in section 2.8.

still giving dramatically different (and more plausible) results in cases like those singled out by the simple argument; my hope is that sketching how this can be done will allow a measure of understanding that counter-examples alone do not.

As a preliminary matter, though, I should manage expectations: I am not aiming to conclusively defend—or even to fully state—this preferred view here. Rather, I am aiming to describe just those features of its structure, in just enough depth, that the exercise will be useful in diagnosing right reason. What follows, then, is a high-level sketch with a very particular purpose.

Such caveats completed, what is the view? It begins with the following anodyne observation about epistemic ‘should’ judgments: in the ordinary course of things, our beliefs about what we *should* believe are not wholly disconnected from what we *do* believe. Suppose, for instance, that I think that hearing eyewitness testimony to that effect that the accused is guilty means that I should believe the accused is guilty. Then, if I hear such testimony, I will (at least sometimes, and *ceteris paribus*) indeed go on to believe that the accused is guilty. And not only does it seem that there is such a connection, but that the connection is no accident: there should, rather, be some explanation for how it is in the nature of our epistemic should judgments that they tend to fit with our believings.

There are, of course, many potential candidate explanations; I am going to pursue just one. Namely, I am going to consider the possibility that epistemic ‘should’ beliefs are intimately tied up with what I’ll call ‘epistemic regulatory states’—states whose aim it is to regulate one’s epistemic behavior, and hence states whose excellence and defect is partially constituted by their so doing. For now I will be silent on what form this linkage takes; are epistemic should judgments themselves regulatory states, or do regulatory states figure in their analysis, or…? I return to that later. For now, the working hypothesis is just: our theory of the epistemic should will wind up in some way citing states which are partially characterized by their regulatory aim. So, what would such a state look like?

Although our target here is *epistemic* regulatory states—because we hope to eventually make our way back to the epistemic 'should'—in answering this question I nonetheless propose we proceed indirectly, by considering a different, *practical* type of regulatory state. Namely, I am going to open the discussion by considering plans for action. When we plan out our morning, a conversation, a trip to the grocery, or whatever else, we engage in an activity that aims to regulate our practical behavior. And it is useful to begin here, rather than directly with epistemic regulation, because the prosaic planning of our day-to-day lives is quite familiar, familiar enough that we have a strong grip on when it does and doesn't 'make sense.' Indeed, I am going to start by outlining what I take to be a highly intuitive constraint on how we may intelligibly plan; I use it to cast light on the constraints that come from plans' goal of regulation. But since they stem from the goal of regulation, rather than the specifics of what is regulated, these will wind up being special cases of constraints on regulatory states *per se.* And so we can thereby leverage the familiar case of practical planning to learn something about regulatory states in general; and this particular lesson, I take it, will pay significant dividends when ported over to the epistemic realm.[23]

There may be many concepts that answer to the ordinary English term 'plan,' but for the purposes of this investigation the general form of 'a plan' will be given by the formula "to φ in

[23] The most fully developed account connecting normative judgments to planning states (or, more cautiously, plan-like states) comes in the work of Allan Gibbard, who gives a full-blown account of all normative language in terms of the expression of a single unified type of practical-epistemic planning state (Gibbard 2003). For more work connecting similar constraints on 'doxastic planning' to generally interesting epistemic conclusions, see Schafer (2014) and Schoenfield (forthcoming b). I chose to use the terminology 'epistemic regulatory state' instead of directly invoking doxastic plans in order to remain neutral on the degree to which the epistemic and practical can be fully assimilated. I want to explicitly avoid, for instance, the suggestion that we can enter into 'doxastic plans' with the same voluntary control we typically enjoy over our more familiar practical plans; whether this is so is beyond the remit of the present concern. The only common feature I require is a regulatory aim; hence, 'regulatory state.'

C”, where that formula maps a ‘condition’ C to a ‘response’ φ. For instance: my plan ‘to eat cereal tomorrow morning’ maps the condition of it being tomorrow morning to the response of eating cereal. I choose this type of plan as my object of interest in order to focus on the relation between conditions to responses; I do so because I think, when we examine it, we see that not just anything goes in terms of how they may be paired up. In particular, I take it that intelligible planning is beholden to the following discriminability constraint: if one believes C and C* are indiscriminable, then planning to φ in C is incompatible with planning to ~φ in C*.[24]

We can illustrate both what this constraint says and also how attractive it is by considering an instance of its violation. So: suppose I believe that tainted milk does not look, taste, or smell any different from untainted milk. In fact, I believe that I cannot detect tainted milk in any way. It’s not just that milk being tainted would make no phenomenal difference, but that I am incapable of differentially responding to it at all, be that by way of a phenomenal intermediary or otherwise. So, for instance, I am not like the chicken-sexer who can tell, without knowing how, what the sex of a newborn chicken is upon examining it. Rather, I believe there is no part of me that can be leveraged to tell the tainted milk from the untainted. But suppose that, despite this, I form both of the following plans: to pour my milk down the drain in the circumstance of it being tainted tomorrow morning; and, to drink my milk in the circumstance of it being untainted tomorrow morning.

This combination is baffling. How can I plan to treat the tainted and untainted milk

[24] The claim here is not that whenever there are some indiscriminable C and C*, and one plans to φ in C, what one ‘really’ plans to do is to φ in C**, where C** is the disjunction of C and C*. It does not hinge on a phenomenalization of the situations for which we plan, cutting away possible error in our recognition of the circumstances until we arrive at a luminous common factor; it is fully consistent with everything said that one can plan to φ in C while having no plans whatsoever for what to do in C*. This is relevant to, e.g., concerns about the absence of luminous conditions, about which more in section 2.8.

differently, while at the same time believing that their difference is undetectable? The discriminability constraint says that I can't—or rather, I can't without implicating myself in some incoherence.

This seems like the right verdict, and the discriminability constraint is correspondingly highly plausible. But what explains it? Presumably, it is not just a brute fact about plans that they cannot be so combined. Rather, there should be something in their nature which makes it so. And, indeed, I think there is a very natural story to be had here, one rooted in plans' regulatory aim.

Forget for a moment about whether the milk is tainted, and return just to my initial plan to eat cereal tomorrow morning. To introduce another term, say that this plan 'triggers' whenever I act on the basis of it. In the ordinary case, and all else being equal, once made this plan will trigger tomorrow morning and in so doing will lead me to eat cereal—which we may suppose is a desirable outcome. And, indeed, this is the point of forming the plan in the first place.

In contrast to that good (and ordinary) case, here are two scenarios that sound bad: tomorrow morning could come, and despite all else remaining equal, I could simply not eat cereal; or, we might imagine that it was not yet tomorrow morning—suppose it is instead the middle of the night—and yet I nonetheless, on the basis of my plan, go to the kitchen and eat cereal. In the first case, my plan doesn't trigger when it was supposed to, and in the second, it triggers when it wasn't. Whatever plans are for, this seems like it can't be it: which is to say that these seem like instances of my plan failing to regulate my behavior appropriately.

Reflection on these sorts of cases leads me to think that plans aim *at the very least* at triggering appropriately; that is to say, at actually triggering under their conditions and not triggering otherwise. Any failure there is a failure *qua* plan. But to simplify the points to come I am going to assume a stronger (though, I think, still quite plausible) aim: to *reliably* trigger in response to their conditions. It is not sufficient for my plan to eat cereal to be satisfied, then, just for it to have actually triggered in the appointed condition. That could happen by freak accident. Plans aim at a form of control, and this control requires the matching of their triggering to their conditions not only in the actual circumstance, but across a range of counterfactual ones as well.[25]

Not only does this aim line up with the cases of success and failure considered, but it looks like it allows us to furnish an easy explanation of the discriminability constraint. Return to the earlier plans for tainted milk: to pour it down the drain if tainted, and to drink it otherwise. If I form both these plans, what does my overall mental state look like, and, in particular, what do I believe about my own plans? I believe that it's possible that when tomorrow morning comes I will act on the basis of one of my plans, and, furthermore, that it's even possible that I will act on the basis of the right one. However, I also believe that if that were to happen, it could only happen as a matter of sheer luck; the right one might trigger, but it would not be a reliable triggering. Since we took reliable triggering to be part and parcel of the regulatory control plans aim at, in believing my plans could not possibly trigger reliably I thereby believe they are in fact defective with respect to their aim. But believing your attitudes are defective in that way is a paradigm form of irrationality; rationality requires doing well by your own lights.[26] Rather, in order to get right with rationality I would need to either change my beliefs about what I can discriminate (which

[25] There are obvious questions to ask here about how to understand the reliability at issue. Do we understand it in terms of safety? In terms of sensitivity? The question strikes me as complicated—and I am actually inclined to say 'neither;' entering into this, though, would take us too far afield.

[26] A complication: preface-style cases involve believing (at least one of) your attitudes is defective with respect to its aim, yet there it seems sensible to maintain the whole set. This point is well taken. A more thorough argument here wouldn't *just* note that one takes one's attitudes to be failing, but rather would give a dominance argument showing that one's current combination of attitudes is always less satisfied than some other fixed-up set, under some suitable measure of degrees of satisfaction. C.f. Joyce's much discussed (1998) argument for probabilism. In general, I am skirting over issues here that I take to live at a level of resolution finer than is appropriate to the present discussion. There are many ways to understand the ultimate explanation of incoherence among attitudes, but I take that the present case is close enough to a paradigm instance that whatever that story is, it will apply without too much fuss.

here seems dubious for substantive reasons) or (as seems right) change my plans so that they instead associate to conditions I *do* think they can reliably track—for instance, by planning to play it safe and decline to drink the milk either way.

So, construing plans as aiming at reliable triggering allows us to furnish a nice explanation for the discriminability constraint. What is interesting, though, is that this explanation, when fully attended to, actually motivates a stronger constraint than the one we started out with.

To wit: in the above explanation, the problem with violations of the discriminability constraint is that they involve holding plans you believe to be defective; the defect you believe they have is that they won't reliably trigger in their conditions. Now, the particular basis on which you believe that defect to be present is that you believe their conditions are indiscriminable to human perception; they won't reliably trigger because it is *physically impossible* for them to do so. But that they can't is just an especially strong reason to believe that they won't. As such, our explanation supports treating the discriminability constraint as a special case of the following *discrimination* constraint: if one thinks one *won't discriminate* between C and C*, then planning to φ in C is incompatible with planning to ~φ in C*.[27]

This constraint covers not only the indiscriminable, but also that which one thinks one will fail to discriminate (even when it is, in fact, discriminable). Here is an example: suppose that I am an amateur thief planning on breaking into a house to rob it. I anticipate that, while I am mid-burgle, I may hear someone coming home. Furthermore, I also know that the noise people make

[27] Compare: suppose I believe 'P' and I also believe 'my belief that P is necessarily false.' This is clearly incoherent. We search for an explanation: we say that the problem is that I take my own belief to be false, and so defective. But notice that the necessity claim in the content of my belief is inessential to this explanation. If my belief that P is necessarily false, then that is an especially good reason to think that it's false. But any other reason would do just as well; that is to say, it wouldn't stop being incoherent if I instead believed P, and also believed my belief that P was merely contingently false.

when they get home indicates the path they’re likely to take through their house, and I know these patterns well enough that I can listen to recordings and reliably say, for instance, whether a particular slamming of doors and running of faucets indicates that a person is going to make dinner or just go to bed. A skilled thief can put this into practice, by staying in the house as long as possible—if they hear someone making dinner, they may linger longer in the bedroom to check the drawers, and etc. I, however, anticipate that I will be so panicked in my first robbery that even though the different sounds I may hear would indicate different things, I will not reliably differentiate them in the heat of the moment.

According to the discrimination constraint, it would be incoherent to hold the trio of: planning to leave if I hear threatening-noises; planning to stay a little longer if I hear unthreatening-noises; and believing that in the heat of the moment I will fail to reliably differentiate the two. This incoherence doesn’t flow from the fact that the noises are going to be indiscriminable to human perception. Rather, we have explicitly said that they *are* discriminable. After all, if you were to play them to me now, in the calm of my office, I could easily sort them into safe and unsafe piles. The problem is that though they are discriminable, when it comes to the specific conditions I’m considering, I don’t think I will discriminate them. It’s not that these plans couldn’t work. It’s that they won’t.

On its own, the verdict that my attitudes in this thieving case are incoherent is not as immediately obvious as the verdict that my attitudes in the tainted milk case are incoherent. But nonetheless, I take it to be still quite plausible, and I take that initial plausibility to be buttressed by the fact that the latter verdict is a corollary of our explanation for that more immediately obvious initial one. So there is a natural arc here, running from the discriminability constraint to its explanation, and from that explanation onward to the discrimination constraint. Once we accept the move from discriminability to discrimination, though, we have arrived at something that we can apply directly to the case which has animated this paper, namely, the case of Sasha the engineer.

Consider: when discussing Sasha, we have described a case—call it the correct case—in which Sasha has done some bit of reasoning correctly, but is then presented with overwhelming (misleading) evidence that she is wrong. She has accurately calculated that the bridge will stand, but now the doctors are telling her that she has brain lesions, etc. We can compare this to the incorrect case, namely the case where Sasha has done some bit of reasoning incorrectly and is then presented with overwhelming (non-misleading) evidence that she is wrong. These cases are discriminable; Sasha has different evidence in each, and that evidence supports different conclusions. Nonetheless, if Sasha forms both the plan to order construction to go forward in the correct case, and the plan to order construction to halt in the incorrect case, those plans will not reliably trigger. This because even if the evidence is discriminable, she is not in a position to anticipate that she will discriminate it.

This is not, or at least need not be, *a priori*. We can allow that there might be someone who took themselves to be like a chicken-sexer for bad calculations, including ones they themselves make. It is just that we do not take ourselves to have this ability, and on the natural filling-in of Sasha's case, neither does she. It would be lovely if merely by planning to differentiate our errors from our successes we could thereby enter into mental states that would reliably do so, but that is not our predicament. As such, if Sasha were to have a set of plans that differed in their recommendations between the correct and the incorrect cases, while maintaining realistic background beliefs about her own efficacy, she would thereby be implicated in a form of rational inconsistency.[28]

The argument above applies to the *practical* regulatory state of planning; Sasha could not coherently plan to order construction forward in the good case and not the bad case. We have not yet said anything at all about any putative *epistemic* regulatory state. However, I submit that if we accept the idea of an epistemic regulatory state, then we will get the very same constraint for the very same reasons. Nothing in the analysis of plans, and the attendant constraints on them, turned on the fact that plans issue in actions. Instead, the analysis turned on the fact that plans aim at reliable triggering, and that when it comes to these conditions Sasha doesn't believe they will. So long as epistemic regulatory states aim at reliable triggering, and so long as Sasha continues to believe that she is not a chicken-sexer for bad arguments, then the same story will apply *mutatis mutandis*. The discriminability constraint will rule out as incoherent her holding of epistemic regulatory states that prescribe differential responses to the good and bad cases, considered as such.

So, if we ask ourselves what coherent set of epistemic regulatory states Sasha might have—states characterized by their aim of regulating belief formation and retention—then we get the answer that none of the possibilities differentiate the good and bad cases, considered as such. But even this is just a conclusion about regulatory states; we still have not yet gotten back to our

[28] Here I'm allowing that a given 'correct' case will have different, discriminable contents from a corresponding 'incorrect' case. Objection: allow that I am right that Sasha would be incoherent if she had separate plans for 'the correct case' and 'the incorrect case' under those descriptions. Still, why can't she form different plans for the correct and incorrect case under the description 'the case with evidence E1' and 'the case with evidence E2?' And so on for *every* pair of correct and incorrect cases? Answer: although this raises important issues, for present purposes it is sufficient to note that evidence admits of infinite variability. As such, one thing our regulatory states may seek to do is regulate our responses to situations lying within the vast sea we have never yet picked out by way of any maximally specific descriptions. Allow that the bridge case is one of these; Sasha does not form in advance, and it would be psychologically impossible to ask she form in advance, any state whose content included a description that specifies the full available evidence. If she is to have a regulatory state that governs her response to the situation at all, it must thereby be at a higher level of generality—and those states have the consistency relations I'm outlining.

ultimate topic, the epistemic ‘should,’ and in order to do so we have to return to the question we initially shelved—what, might we suppose, is the relationship between epistemic ‘should’ beliefs and epistemic regulatory states?

I shelved this question partially because I do not take it to be essential to the thrust of argument here. Rather, I take it that there are many answers each equivalently good for my purposes, where which one opts for is a matter of broader theoretical taste independent of the present application. But to get a sense of things, we may very loosely canvass some. So: one very ambitious answer is expressivist in character. On this line of thought, claims about what a person should believe in a particular circumstance just *express* a hypothetical planning state with suitably arranged contents. More modest answers are also available. For instance, regulatory states may feature in an analysis of the epistemic should that also contains some other familiar components, like an ‘ideal advisor’ or a ‘constructive procedure’: a person should believe something in a circumstance iff a suitable ideal advisor would want them to have a regulatory state matching that belief to that circumstance, or a person should believe something in a circumstance iff having a regulatory state which matched that belief to that circumstance is the output of some suitably specified procedure on their attitudes.[29] Perhaps yet most modestly, one could think that expressibility in consistent regulatory states is just one among many independent substantive constraints on what the epistemic ‘should’ facts are; an ought-implies-can-regulate principle, if you would. I take it that on any of these strategies for connecting plans to ‘should’ judgments, there will be a very natural story leading from the consistency constraints we’ve outlined to the further conclusion that one shouldn’t believe what the right reason view recommends.

[29] See Schoenfield’s (forthcoming a) for versions of this strategy. She considers both what plans perfectly rational but otherwise ignorant advisor would want you to *actually* follow, and which she would want you to *make*; she holds each connects to rationality in an interesting and distinct way.

I take it that the key step, then, is not how particularly one connects epistemic regulatory states and epistemic 'should' beliefs. There are plenty of ways to carry that out. The key step, rather, is allowing that there is something like an 'epistemic regulatory state' in the first place, and being willing to give it some important role in the analysis of the epistemic 'should.' At the start we motivated the introduction of such states by observing that epistemic should judgments *do* tend to match with epistemic behavior, and then positing some connection to distinctly regulatory states as an explanation. This is alright so far as it goes, but it only takes us so far. So, I will take a moment now to highlight two additional virtues of the present account, and, by extension, two additional virtues of the decision to allowing regulatory states an important role in our understanding of the epistemic 'should.'

The first is that the discriminability and discrimination constraints on prosaic plans for action are both minimal and independently well-motivated; if the fraught behavior of epistemic 'should' judgements in calibrating cases can be satisfactorily explained by assimilation to them, then we will thereby have reaped great rewards by appeal to only very sparse resources.[30]

The second is perhaps more salient in the present context. Recall that when we listed motivations for the right reason view I allowed that those motives were genuine—but held that they did not support right reason *per se*. Rather, I suggested that they supported the idea that we need *some* 'rigid' epistemic concept in our toolbox—one that may be preserved under competent deduction, one that may be *a priori* and indefeasible, and so on. We are now in a position to see

[30] An example of these resources in action: in Titelbaum's argument for an evaluatively-focused version of the right reason view in his (2015). He challenges his interlocutors to provide a reasonable theoretical picture on which akratic combinations of beliefs are ruled out, but which does not thereby wind up entailing that all mistakes about rationality—not just mistakes about what it demands in one's current circumstances—are also ruled out (Titelbaum 2015: 21). As far as I can tell, there is a straightforward answer on the present picture, as akratic combinations will involve guaranteed failure in one's regulatory states but mistakes about what is rationally required in other circumstances will not.

how such a concept could play an interesting and important role without rising to the level of determining the epistemic ‘should.’ For notice what we have made a point *not* to say. We did not say that Sasha was in *the same total epistemic position* regardless of whether her reasoning was or was not actually correct, and that her being in that same total position was why she could not coherently plan to believe different things. We have allowed that her total epistemic position is different; after all, the cases are discriminable. Rather, what we have been developing is an argument for why *even if they are discriminable*, it may be that Sasha nonetheless can’t consistently adopt regulatory states that aim to discriminate between them. This because even if they are discriminable, she may not believe those states would reliably discriminate them.

The space of options one believes to be discriminable may be very different from the space of options one believes one will discriminate; and, concordantly, each space may be crucial to a different normative concept. In the context of the present debate this importantly allows us to concede, if we are impressed by the arguments for right reason, that evidence goes the way of discriminability; we may still hold that there’s another normative concept that goes the way of discrimination. We can thereby allow that the right reason view is fully correct in its assessment of the weight of one’s evidence; the weight of one’s evidence is a matter of one’s actual justification and that’s it. But nonetheless, we have still found an important role for calibrating features to play. In the cases we’ve discussed, calibrating features are hooks one’s regulatory states can grab onto. One does not believe one’s regulatory states will reliably trigger when attempting to hook up to the weight of the evidence, specified as such, for just the same reason one does not think they will reliably trigger when attempting to hook up to the correct and incorrect cases, specified as such. But there is no reason to think one’s regulatory states will be unable to reliably trigger in condition like “a doctor tells me my reasoning is compromised”—

believing one’s regulatory states to have the power to reliable trigger in that circumstance is not on a par with believing oneself to be a chicken-sexer for arguments. And so calibrating features present a set of valid targets, targets which are attractive in the context of managing one’s belief formation and retention. And so, even given that they do not correspond directly to the weight of one’s evidence, they may still be crucial to the epistemic ‘should,’ a concept we have proposed to understand in terms of that activity.

This is the sense, then, in which I take it that the present view not only presents an alternative to right reason, but presents an alternative that grows out of appreciation of it. The present view has the right structure to allow one to be impressed by many of the arguments in its favor, and yet resist the idea that it correctly describes the epistemic ‘should.’

Still, for all the lovely things I may say here about it, I do not pretend to have established the correctness of my preferred view. There are serious problems to be worked out; that is why the paper starts from the simple argument against right reason, tries to show that it is inescapable, and only then goes on to offer (more speculative) diagnoses and alternatives.

It is also worth noting that there are other alternatives, too—which is to reiterate the claim at the beginning of this section, namely, that the central argument against right reason does not hinge on accepting the positive sketch here. So to readers who find it particularly unappealing, I say: we can recall, for instance, the virtue-theoretic view on which reasonability, as exemplified in the reasonable person, is the normative standard most directly relevant to belief and action; such a view makes no obvious appeal to regulatory states or anything like. These views are perhaps worthy of more investigation. The important thing, here, is just what such a view is not: namely, it is not a version of the right reason view.

2.8 LUMINOSITY AND NORMATIVE DIVERGENCE

In this paper, I have identified a recipe for cases where the right reason view appears to yield the wrong result. My argument relies heavily on an intuition about cases: Sasha ought not order the construction to go forward. I haven't just rested there—I have also gestured at a theory that would neatly account for those intuitions. But there is still a significant strain of current philosophical thought according to which the whole exercise was pointless: on this view, there is simply no point in trying to accommodate the intuitions I started with, because they are essentially flawed. I suspect this conviction hangs behind much support for the right reason view and so I am under some obligation to address it.

I have identified cases in which the truth of the right reason view would lead to what we might call, following Hawthorne and Srinivasan, *normative divergence*: here they are cases where a person is doing things our putative norm sanctions as right but yet the person is intuitively blameworthy (Hawthorne and Srinivasan 2013).[31] I have taken these cases to be ones where the right reason view cannot avoid getting the wrong answers, and so I have taken these cases to thereby refute it. But, the response goes, even if the right reason view has the property of generating such cases, this is not a strike against it. Given that Williamson has demonstrated that there are no 'luminous' conditions, it follows that any norm whatsoever will lead to cases of normative divergence. So there is no *special* sin attached to right reason; this is just the sort of thing we must learn to live with.

To fill in a bit: how is anti-luminosity supposed to generate normative divergence? Anti-

[31] For them, normative divergence is generally when a norm's recommendation comes apart from our judgments of blame, but blameworthy right-doing is the type relevant here.

luminosity teaches us that for any non-trivial condition there are cases where it obtains, yet we are not in a position to know it obtains.[32] That a non-trivial norm applies to our current situation is itself a non-trivial condition, and so, for any such norm there will be situations in which it applies to us and yet we are not in a position to know it applies. If we fill in the case in the right way, we can then get ourselves to feel the intuition that it would be wrong for us to proceed in accordance with that norm, and so we get a case of normative divergence. But none of this has turned on the content of the norm; these considerations will apply to all norms.

Clearly, whether this argument even gets off the ground hinges on what, if anything, Williamson has really taught us. But adjudicating that issue is beyond the scope of this paper.[33] However, even granting *both* that there are no luminous non-trivial conditions *and* that this feature means that all non-trivial norms will yield cases of normative divergence, this is not yet adequate as a response to the simple argument. And the reason is straightforward. Even if there must be some cases of normative divergence, that doesn't mean we have to count Sasha's among them.[34]

Consider the sketch of the regulatory-state view offered in section 2.7. Nowhere did it appeal to luminosity; it is consistent with luminosity-failure. If the argument we are currently considering is correct, then there must be some luminosity failures, and they must lead to some cases of normative divergence, for *whatever* particular specification of the regulatory-state view

[32] Williamson has given a number of related arguments for a number of related 'anti-luminosity' theses; the particular thesis given here is the conclusion of his canonical (2000) argument. The common premise essential to all these arguments is the claim that for any case where a non-trivial condition obtains, there is a chain of pairwise indiscriminable intermediate cases linking it to a case where that condition clearly doesn't obtain.

[33] There is a significant literature on what precisely Williamson has taught us. For doubters, see e.g. Berker (2008), Cohen (2010b), Fumerton (2009), and Smithies (2012). Running defense, see Srinivasan (2013).

[34] For similar reasons, my argument in this paper is fully compatible with Lasonen-Aarnio's (2014). If Lasonen-Aarnio's argument maximally succeeds, it shows that our rules cannot be defeasible 'all the way up.' But it is compatible with everything here that there are *some* cases where actual justifications are indefeasible in light of calibrating features. What is under attack is rather the claim that *all* cases are such that actual justifications are indefeasible in light of calibrating features.

we endorse. However, the fact that we must allow *some* normative divergence is no argument for allowing it everywhere, especially when it is grossly offensive to our pre-theoretic sensibilities. Why not prefer a theory that, though it allows for such divergence somewhere, avoids it in the treatment of Sasha's case?

The implicit premise being appealed to here by the defender of right reason is that normative divergence is the sort of thing where, once we allow it, there is no additional cost to allowing an infinite amount of it. And this, I think, would be a perfectly reasonable outlook to take under certain conditions. For instance, it would be reasonable if we thought that we had a general story that in any given case could reconcile normative divergence *at some level* with our overall judgments: if we thought, for instance, that the epistemic 'should' could go the way of divergence, but that we could always appeal to risk-avoidance, institutional obligations, good cognitive habits, reasonable dispositions, or something of the like to smooth over the apparently jaw-dropping practical consequences. With extra machinery to make right of the world once again, we could indeed let normative divergence roam free. But it has been the task of this paper to argue that no such stories work.

When we are in a realm where there is no comforting further story to tell about normative divergence, it is entirely worthwhile to minimize its occurrence, especially with regard to cases we feel strongly about. And so, even under quite generous assumptions about the force of anti-luminosity arguments, accommodating the desired result in Sasha's case remains a fully adequate motivation for rejecting the right reason view.

3.0 AGAINST INTERACTIONISM

3.1 INTRODUCTION

We perform all sorts of reasoning tasks: we try to divine who will win the election, whether we'll save more money if we buy or rent, where to look for our missing keys. One of the things we can reason about is our reasoning itself. We can raise concerns about whether our reasoning was correct, and we can then come to doubt or modify our conclusion on that basis. We can raise those concerns even in cases where one *has in fact* reasoned perfectly correctly, but yet still manages to acquire *very strong reason* to think one has made a mistake.

For instance consider *Dining Table:* I am at dinner with a group of friends. One proposes: "Let's play a game wherein I write out a formula in sentential logic and then we determine whether it's true on all assignments." She writes something out and we all pause to think it through. I reason informally—'because the central connective is a conditional, for there to be a false assignment I must first make sure the antecedent comes out true...' and so on. I conclude that the formula is true on all assignments. Suppose also that I am right. Nevertheless, much to my consternation, when I share my answer I see that my companions have unanimously arrived at the opposite result. I also realize that I have had too much to drink, and, furthermore, I know that my friends are better at logic than I am. So I have a superabundance of reasons doubt myself. After all, on reflection it seems very hubristic to suppose I was the one to get it right and that they all got it wrong. That would be something of a miracle (—although of course, in this case, that miracle is exactly what happened).

In this case, the answer I initially arrived at was right, despite the subsequent indications otherwise. What is the significance of that rightness, when it comes to deciding the question of what I ought to believe at the end of the day? I will argue that there are only two good answers. Either the rightness of my initial answer 1) is the *only* thing that matters, or 2) it doesn't matter *at all*. My goal, then, is to argue against the moderate proposal on which whether I initially got it right matters somewhat, but so too do the later reasons I got to doubt myself. On my reading, such otherwise diverse figures as Kelly (2010), (2013), Setiya (2012), Lackey (2008a), (2008b), Sliwa and Horowitz (2015), and Christensen (forthcoming) all have views that are moderate in this way. So the upshot of my argument will be that though they are quite different in their details, all these authors' views can nonetheless be ruled out on the basis of a shared structural feature. So I will argue that the moderation for which they aim is not on offer.

Now the characterization in terms of 'the rightness of my answer' and 'mattering,' is, I think, intuitive, but can do with some unpacking. It will be useful to introduce some terminology. We can begin with the (now fairly standard) distinction between *first order* and *higher order* evidence.[35] First order evidence is, well, whatever evidence we directly reason from in any given reasoning task. Since the dining table case involves a logical reasoning task, what we take the first order evidence to be there will depend on the particular epistemology of logic we accept. But whatever that is, it will presumably give us some evidential route by which we can come to justified logical beliefs. However that particular story happens to go, it will be highly plausible that we can later come to have reasons to gain or lose confidence that we have successfully met its specifications (whatever they were). Higher order evidence consists in reasons of that latter sort: it is evidence that affects our confidence in our final answers not directly, but rather *only* by way

[35] See (Christensen 2010) for an exploration of the distinctive characteristics of higher order evidence.

of intermediately affecting our confidence in the correctness of our first-order reasoning. Here the fact that all of my companions have come to a contrary result affects my confidence *by* making me unsure of whether my initial reasoning was correct. After all, someone must have made a mistake and all indication seem to that person being me.

Sometimes the distinction between first- and higher-order evidence is messy. In the present case, though, it isn't. Recall the reasoning that's in question: 'the central connective is … so in order to make the whole false… but that can only happen if…" and so on. Again, there is presumably some true epistemology of logic that will tell us how exactly this reasoning is to be captured in first-order terms. The important thing, though, is to notice the kinds of consideration that do *not* figure in episodes of reasoning like the one given above. That reasoning features no claims about *my formal aptitude*, *how much I've had to drink*, *what someone else at the table thinks*, or anything of the sort. Indeed, those facts are not even candidates to appear in the logical form of reasoning that lead to my (provisional) answer. And, given that that is so, if those considerations are to bear on whether the formula is a tautology it seems that they must do so indirectly, by way of bearing on my confidence in my reasoning: how could the apparently unrelated fact that *I'm drunk* bear on *the formula being a tautology* except by way of the connecting fact that I have just (drunkenly) done some reasoning which is now under scrutiny? So here we have a nice sharp distinction here between first-order reasoning and higher-order reasoning, and concomitantly between the reasoning that tells on the answer directly and the reasoning that tells on the answer only indirectly, by way of mediately influencing our level of confidence in that more direct line of reasoning.

With that distinction in hand, we can now give a gloss on what I meant by "the initial right answer." An answer is initially right just in case it is possible to arrive at by successfully completing the reasoning task at hand. In this case, there is exactly one such answer.[36] The formula is, after all, a tautological schema, and so good logical reasoning can only ever lead to the conclusion that it is; reaching the conclusion that it isn't requires committing some error. Since we have conceptualized success at our reasoning task in terms of responding to some body of first order evidence (whatever it is), we can then helpfully go on to speak of these right answers as those that are "supported by the first order evidence," and, similarly, we can also speak generically of "the support of the first order evidence" whenever we are interested in which those initial right answers are. So the question of *whether it matters what the initial right answer is* can be glossed for our purposes as *whether it matters what the first order evidence supports.*

But what, then, if this last concept, this concept of *mattering?* Of course, things matter in all sorts of different ways. In the present context, though, I am interested in a particular way of mattering. I am interested in which sorts of things directly determine what I ought to believe. For a parallel, think of the evidentialist. Evidentialists allow that all sorts of factors may indirectly influence what you should believe. They maintain, though, that they only do so by mediately affecting the evidence you possess, and that it is only ever your evidence that directly determines what you ought to believe. In general, when we give normative epistemic theories we single out some factors for special attention: perhaps our evidence, perhaps the reliability of our cognitive processes, perhaps our knowledge plus the intrinsic credibility of hypotheses, etc.; we then say that settling those factors thereby settles what you ought to believe, regardless of what else may

[36] That there is a unique right answer simplifies the case. I believe that the arguments to follow can equally well go through if that assumption is relaxed, though I will not address that here. For further work on the relation between first-order uniqueness assumptions and the significance of higher-order evidence see Kelly's (2010) and Christensen's (2011).

be so. Those are the factors that “matter,” in the sense that I am deploying it here. I am interested in what matters in settling what one ought to believe, when one is in a problem case.

In this paper, I am going to argue that the first order evidence and higher order evidence cannot both matter at the same time. One might matter, and the other might matter, but it’s not the case that they both do. Of course, if this is so it leads us to wonder which matters, and for what. I will, by the by, say some small piece about how I think we ought to navigate that choice, but nonetheless the primary purpose of this paper is to force the issue.

3.2 INTERACTIONISM

My characterization of interactionism is rather abstract: the label is chosen to be broadly encompassing, and so must prescind from the (varied, interesting) details of the otherwise heterogeneous views falling under it. But though running the argument at this thousand foot view pays nice dividends in increasing its scope, it will nonetheless be helpful to begin with an orienting example. For that purpose, I introduce a simple example of an interactionist view.[37]

So: let us consider what I’ll call “the weighing view.” The weighing view is an interactionist view, and so by definition it holds that both the first order evidence and the higher order evidence matter. But, additionally, it also tells us something about *how* they matter. It tells

[37] This view is not that dissimilar from the views presented in (Kelly 2010) or (Lackey 2008a) and (Lackey 2008b), though it is both simpler and correspondingly less plausible. For instance, Kelly does not commit to the claim that first and higher order evidence *always* weigh against each other, but holds just that they sometimes do, and he includes a more complicated explanation of how this weighing occurs (namely, by way of competing downward and upward ‘pushes’ that go through an intermediate epistemic proposition). And Lackey doesn’t put her view in terms of weights—or pushes—as a conceptual model, but just takes her inputs as the degree of first-order justification and the strength of the higher order evidence. Though from her examples, it seems clear that her view often generates the same results as the weighing view, it needn’t be by the same procedure. Fully teasing these differences out, though interesting, wouldn’t do much to profit the base argument.

us that the way each body of evidence matters is by placing some weight on the conclusion that it supports, weights that must be balanced against each other to come to a final conclusion. In so advising, the weighing view analogizes the contribution of each type of evidence to the contributions of two more mundane separate sources. So, for instance, imagine that we are climate scientists and we find that the pollen core data tells in favor of some model, but the satellite data tells against it. When we have both of these bodies of evidence together, we may end up weighing them against each other: if the pollen core data is stronger, it outweighs the satellite data, if they are about the same, perhaps they balance, and so on. And on the weighing view, this is just the same thing that happens when one is in possession of both some first-order and some higher order evidence. Each type of evidence provides a weight in favor of some conclusion, and in balancing those weights we arrive at the all things considered right answer.

So, in the dining table case, we could apply the weighing view as follows: the first order evidence (however we conceive of it) supports the conclusion that the formula is a tautological schema, since, after all, it is. The higher order evidence, though, supports the conclusion that it isn’t. The others getting the contrary result and my various reasoning impairments all agree in pointing toward my having gone wrong, and so toward the other answer being right. So we have two weights pushing in opposite directions. When those weights are balanced against each other, the result will be that ultimate confidence I ought to have should lie somewhere in the middle. Where exactly that balancing point lies will depend on the exact weights we assign, which will in turn depend on working out more specific theses about how the weights of bodies of first- and higher order evidence are independently determined. But although that is left open, there is no mystery about how they then go on to combine once so determined. They combine just the way that any two mutually rebutting bodies of evidence do.

So, that is one simple sort of interactionism. The first order and higher order evidence both matter in the dining table case in just the same way that the pollen core and satellite data both matter in deciding a climate model. The way they matter is that each weighs against the other and the result of that weighing determines an all things considered verdict. The first order evidence and higher order evidence are inputs to a function—here, weighing—which then tells us, on their basis, the correct doxastic state to hold. Interactionism, by contrast, is the class of *all* views which have both the first and higher order evidence as inputs at the same time, regardless of what they happen to go on to do with them in the process of determining an answer.

We can expand on this for a moment. For some purposes, we can think of epistemic theories as describing functions, functions that take some epistemically relevant features as inputs and then map those inputs onto correct doxastic states. Theories can disagree both on what features ought to be inputs and on what procedure ought to transform them into outputs. For instance, classical subjective and objective Bayesians disagree only over inputs, namely, whether the inputs to any given learning task should include the agent's subjectively held prior or instead an objectively correct one. They agree, by contrast, on conditionalization as the subsequent procedure for mapping those inputs to correct resulting doxastic state. By contrast, a crude process reliabilists departs from them on both counts, thinking that the relevant inputs are instead belief forming processes, their issuances, and their associated reliabilities, and that the subsequent procedure for obtaining the correct doxastic states is one that maps the issuances of the high reliability processes to justified resulting beliefs.

As above, 'interactionism' is a claim about inputs and not procedures. Again, a view counts as interactionist iff it holds that there exist cases where the correct doxastic state is determined by a function of *both* the first and higher order evidence. Although any interesting interactionist view will presumably have something more to say about how the first order evidence and the higher order evidence are supposed to come together to determine a correct subsequent doxastic state—e.g. by weighing—those features will not be crucial to the argument of this paper.

Now, this language of procedures, functions, and so on can be misleading. It's important to be maximally clear that nothing here commits one to the further claim that conforming with an epistemic view requires *consciously calculating* some function. We need not over intellectualize, and the example of reliabilism above should illustrate this. After all, a key tenet of reliabilism is that it is reliability, not our thoughts about our reliability, which is epistemically relevant at least in the first instance. And so, similarly, when I classify interactionist views as those that take both the first and higher order evidence as inputs, I am not thereby implying that interactionists ask agents to ever consciously compute either of those things. Indeed, with respect to the weighing view, we are perfectly free to imagine that both the evidential assessments and the weighing of each against the other typically takes place sub-personally. For all that, it would still count as a species of interactionism.

Nonetheless, although complying with an epistemic view does not, as a conceptual matter, imply having that view and consciously applying it to one's situation, it will nonetheless often be true that we do happen to have some conscious knowledge of our epistemic view and of how our performances could—or could fail to—execute it. So again, a generic reliabilism does not require that I have some antecedent knowledge of e.g. my visual reliability, knowledge which I then put into practice by trusting my sight. I never need to think about the reliability of my vision at all. Nonetheless, if I *do* think about it I can conclude a few things. As a human adult with ordinary introspective capacities, I'll be able to ascertain and affirm propositions like "I'm treating the outputs of my visual processes as reliable" and "on reliabilism, the beliefs I'm forming on the basis of my visual processes are justified only if the way I’m treating my visual processes—as reliable—is indeed accurate.”

Now, to foreshadow, my critique of interactionism will make central use of such self-assessments. So right at the outset I want to be clear about what that implies for my target. It does not imply that my target—interactionist views, as I conceive them—are limited to those on which such limited self-knowledge is conceptually entailed by the view itself. My target includes views on which that is not true. Rather, my thought is that there is a problem with interactionism that we see manifested when we turn our attention to agents who, as a purely contingent matter, do happen to have the sort of self-knowledge typically possessed by adult humans. I will of course say more about both the ostensible problem and the sort of self-knowledge that brings it to light when I give the actual argument, but here the point is just to be very clear about how little I am building into the characterization of interactionism itself. Interactionist views are *just* those that take both types of evidence as inputs.

For all that I am *not* assuming, though, I do have to offer something of a caveat. In my central example, the one to which I will keep returning, the agent makes some initial first order judgment and then there is a temporal separation before the higher order reasons for doubt come to light. First, I think through the formula and judge it is a tautology; second, I discover that all my companions got the contrary result and etc. But nothing about either interactionism or the characterizations of first and higher order evidence obviously implies such a temporal ordering. So it may look like I am fitting my discussion exclusively to a special case and hence drawing potentially misleading verdicts. This criticism is not entirely unfair. Indeed, I think there are knotty philosophical issues involved in generalizing away from the first-order-judgment-first higher-order-evidence-later ordering I’ll be working with. But I will not enter into them. Rather, for reasons of both stylistic ease and conceptual simplification, I will continue to speak interchangeably of one's "initial verdict" and one's judgment of the first order evidence, and similarly of one's "subsequent verdict" and one's all things considered judgment after the addition of higher order evidence. A fully adequate general treatment of higher order evidence wouldn't do this. Still, in my defense, for the main purpose here—critiquing interactionism—I don't need to offer anything like a fully adequate general treatment. Rather, even if they are not the only kind of case that exists, so long as at least some cases with this temporal structure *do* exist it's then fair game in critiquing interactionism to show how it flounders when applied to them. So I raise this issue to bracket it.

Now at this point I have said a fair amount about what interactionism is. I have not, however, said much about why it is interesting or important. Why bother to critique it? Here's a reason: there is a plausible argument that *only* interactionists can get all the intuitive cases right, because getting all the cases right requires invoking sensitivity to each type of evidence in its proper place. And so that is the argument to which I now turn.

3.3 AN ARGUMENT FOR INTERACTIONISM

Interactionists views take both the first order and the higher order evidence to matter. We can motivate the view that both matter by arguing that each matters. Here's the slightly indirect strategy I'll pursue for doing so. For each type of evidence, I will first consider an argument that it *doesn't* matter. Then I'll consider a rejoinder for each which purports to show that there are some cases we can *only* get right if the relevant type of evidence matters: our theory needs to be able to 'see' that type of evidence in order to give a fully adequate set of verdicts. The result will be that we end up with an argument that each type of evidence needs to matter in at least some cases. And once we get that conclusion, the most natural resulting view will be one on which both matter in a wide range of cases. So we end up as interactionists.

First, let's start with the argument that higher order evidence doesn't matter. How might one talk oneself into that conclusion? Here's one line of thought: higher order evidence isn't *about* the right sort of thing. In the dining table case, my higher order evidence consists in a bunch of facts about what my companions judged, whether I've been drinking, and so on and so forth. But these facts don't bear on whether some formula is a tautology. How could a bunch of empirical dreck ever be evidence of the logical status of some formula? After all, if we say that these facts are evidence that the formula isn't a tautology, are we also going to say that they are evidence that some non-standard logic is correct? And how would this evidential support relation work—not by evidence supporting that which it entails, given that throughout my evidence has always (vacuously) entails that the formula is a tautology. These questions are hideous and lack good answers.[38] Better instead to maintain that both before and after I receive this higher order evidence, I am already in possession of the strongest possible evidence that the formula is a tautology, i.e. a sound deductive argument to that effect.

Now, perhaps the higher order evidence *is* evidence for some distinct yet closely related propositions. And so perhaps it could matter when it comes to deciding those related propositions. In this case, it may be excellent evidence *that my companions are bad at logic*, or perhaps even *that my companions are playing a trick on me.* In a more concessive and also somewhat more

[38] Weatherson, in his (ms.) presses this objection; similar considerations are raised in Kelly's seminal (2005). Titelbaum's (2015) argument to the same conclusion—the conclusion that higher order evidence doesn't matter—takes a different path, but also essentially relies on the *a priori* character of judgments about what the evidence supports.

complicated vein, it might be evidence for the non-logical, epistemic proposition *that I am not rational in concluding that the formula is a tautology*. But whatever it is evidence for, it just isn't evidence that the formula isn't a tautology. And so when it comes to deciding whether the formula is a tautology, as opposed to deciding any of those other related questions, the higher order evidence doesn't matter.

However, the interactionist can complain that whatever plausibility this view has in relatively ordinary cases, it falls apart when we consider its application as taken to the extreme. Consider what happens when we keep adding more and more undercutting higher order evidence. At a certain point we'll reach what I'll refer to as "overwhelming defeat."[39] Suppose that we are in a case similar to the dining table case, but instead of a handful of relatively ordinary dinner companions, this time it is *a million* highly trained logicians and all have unanimously and independently arrived at the contrary answer. It is simply mind-boggling to suppose that even upon learning this I should still be just as confident, and not one whit less, in holding on to my original conclusion—provided, of course, that I was initially right. Given that overwhelming defeat is a genuine phenomenon, our theory needs to be able to see the higher order evidence in order to get the right result in those cases where it piles up so dramatically. So the higher order evidence does have to matter after all.

So that is an argument for why the higher order evidence has to matter. Why does the first order evidence have to matter?

As before, we start by asking how one might convince oneself that the first order evidence *doesn't* matter. Here's one natural line of thought: it would be all well and good if I could simply

[39] Lackey in her (2008a) dubs a variant of this the "one to many" problem; in Titelbaum's (2015) he calls it the "crowdsourcing" objection. In general, this objection has been widely noticed and discussed. See chapter 1 for a deeper treatment of the various attempts to defend against it.

believe what the first order evidence supports. But the whole point is that I can't: that in cases like the dining table case I'm uncertain of what it supports and trying to make a determination of its support on the basis of situationally relevant clues. An appearance/reality distinction has opened up between my initial judgment and the first order evidence itself. In trying to correct for this gap it is senseless to appeal to the first order evidence itself, just as senseless as if I told someone who was unsure about the shimmering on the horizon that they ought to believe there is an oasis just in case there actually is one. The support of the first order evidence is the object of my investigation, and so not something which I can simply appeal to within that selfsame investigation.

Of course, the fact that the (intellectual) reality itself is not directly available doesn't mean that the appearance of intellectual reality isn't. And the parallel with the visual case here is instructive. When we gauge whether there is an oasis on the horizon, we do so using information about how likely an appearance like the one we're facing is to be correct given the background conditions in which it was made: we ask about how reliable we expect such a shimmering to be given that we are confronting that appearance as generated in bright light near noon, while it is very hot, in a sand desert, and so on. But of course we do not describe the relevant background conditions in such a way as to presuppose that there is (or is not) an oasis. I don't ask how likely the shimmering seeming I'm experiencing is to indicate that there is an oasis given that it was generated by bright light *bouncing off an oasis* near noon, and so on. And similarly, when we try to correct for a divergence between intellectual appearances and reality, we will ask ourselves how likely a judgment like the one we made was to be correct, given the circumstances. But again we will be careful to describe the circumstances in ways that do not prejudge the question one way or the other. We ask, e.g., how likely is it that the formula is a tautology, given that it seemed to me like it was, and where that seeming occurred under conditions where I was tipsy, contradicted by friends, whatever. We don't ask how likely it is that the formula is a tautology, given that it seemed to me like it was, and that seeming occurred under conditions of *having validly deduced it* while tipsy, etc.

Getting into this frame of mind is one way to talk yourself into thinking the first order evidence doesn't matter. The only things it is kosher for our epistemic theory to directly appeal to are the content of the initial judgment (right or wrong) and then higher order evidence about its reliability—as conceived of in a way that is independent of, and hence brackets out, the first order evidence in the case at hand. So first order evidence gets to participate indirectly insofar as it may be a cause of our initial judgments, but only indirectly: once the content of your initial judgment is fixed, the first order evidence plays no further role in determining the correct final judgment.

Now, there is certainly more than one objection to this type of picture. But I'm going to pass over some of the more theoretical ones in favor of addressing another essentially extensional point.[40] The objection, as before, runs that there are certain cases that you simply *can't* get right without appealing to the disposition of the first order evidence. And so, and again, the theory needs to be able to see the first order evidence in order to render satisfactory verdicts on these cases. If this objection holds water, it has the benefit of working quite generally against any view on which the first order evidence doesn't matter, not just on views which rely on the (potentially problematic) picture given above.

[40] For expressions of theoretical unease about the assimilation of intellectual 'appearances' to visual appearances, see especially (White 2010) and (Kelly 2010). By contrast, the extensional objection that I am going to focus on has its most direct genesis in Kelly's (2010) consideration of the single peer case, but also is anticipated by complaints in (Lackey 2008a); it has what I take to be its most systematic development in (Schoenfield 2014). It is worth noting that it is not the only extensional objection. There has also been a degree of general skepticism that a theory which brackets the first order evidence can capture ordinary cases; see (Kelly 2010) and (Christensen 2011) for a locus classicus of this debate, and (Lord 2013) for criticism along these lines. My view is that the general skepticism typically rests on either misapplying the bracketing procedure to thereby get a bad result, or, alternately, a correct application of the bracketing but one that then produces a result that is only tendentiously bad. But this discussion descends into the weeds rather quickly.

So, how does it go? Here’s the thought: sometimes we make rational mistakes even with respect to questions we are generally competent to answer and in cases where conditions are good. For instance, suppose I am usually pretty good at forecasting whether a romantic suitor is going to want a second date. But on this particular instance I happen to make a mistake: I conclude that they’ll want a second date even though the first-order considerations—first order considerations I ordinarily correctly respond to—say otherwise. So I wind up irrationally believing something I shouldn’t have. But, the objection goes, this result cannot be recovered without allowing the first order evidence to matter.

Why is that? It’s because all the higher order considerations are favorable—I am generally a good judge, there were no distorting or unfavorable features of the circumstances—and similarly, any corresponding estimate of my reliability at judging things in these circumstances would be good. So any account which is only sensitive to the higher order evidence must correspondingly say that I *should* be confident in my conclusion. After all, all the higher order evidence is favorable, and so if that is the only thing my theory sees then my theory only sees good things! But the obvious verdict about the case was supposed to be the opposite, namely, that I had made a mistake, and that as a result I am irrationally overconfident of my conclusion. To say otherwise would allow laundering bad beliefs into good ones just by way of reflection on one’s general competence. Given how often we have information indicating our general competence, this, in turn, would be far too close to obliterating the category of irrational belief entirely. To avoid this consequence, the first order evidence has to matter.

Together, then, those are motivations to be an interactionist. If our theory ignores higher order evidence, it can’t capture the fact of overwhelming defeat. But if our theory ignores first order evidence, it will illicitly launder away what should be counted as genuine rational failings. So we can't ignore either and both have to matter.

Now, astute readers will notice that this conclusions still falls short of establishing interactionism proper. Why? Because although upholding these motivations requires us to take the first order evidence to matter in some cases, and similarly requires us to take the higher order evidence to matter in some cases, it does not thereby require us to admit that they ever both matter in *the same* case. Perhaps sometimes it is the one, sometimes it is the other, but never is it both at once—where that it sometimes is both at once is the claim that defines interactionism.

To which I say: fair enough. I will note, though, that there are *prima facie* difficulties for such a switching view. Notably, the motivations we considered for sometimes ignoring each type of evidence—higher order evidence is about the wrong thing, first order evidence isn't available in the right way—are motivations for *always* ignoring them. If we were to have a view where we ignored either on a case-by-case basis, we could not motivate it by appealing to either of those thoughts. We would need a new set of motivations. And furthermore, such a switching view looks *prima facie* less elegant than a view on which both first and higher order evidence matter across the board.[41]

We can see, for instance, how *easily* the weighing view—a view on which both matter across the board—can account for the phenomena we have raised in this section. The weighing view trivially accommodates overwhelming defeat. As one gets more and more higher order evidence that one has made a mistake, the weight of that evidence thereby continually increases;

[41] Nonetheless, for purposes of full disclosure, I will say that I do think that some versions of the switching view are indeed promising. Their place in this dialectic, though, is a bit tricky—because what I take to be the best variations on the switching view do not preserve the motivations given here. That is to say, they do not respect overwhelming defeat and prohibit laundering. So they do not represent an alternative to interactionism *with respect to maintaining those motivations.*

as its weight eventually approaches stratospheric levels it comes to (almost) entirely swamp the weight of the first order evidence. So overwhelming defeat is indeed capable of pushing one's confidence arbitrarily low. And the weighing view also has an easy time avoiding laundering. When one initially botches one's assessment of the first order evidence, one thereby winds up putting that weight in the wrong place. On ordinary assumptions about how the weighing works (assumptions that maintain the analogy to physical weight), getting the first order weight wrong will mean one continues to get the ultimate weight wrong. There is no danger of some easily available higher order evidence transmuting all your errors into successes. And so we see our sample interactionist view indeed not only can accommodate, but does particularly well accommodating the phenomena that motivate it.

So those phenomena then present us with some reason to be an interactionist. Interactionists, the thought goes, can account for features of epistemic life which their more myopic cousins cannot.

But anyway, enough of this. After all, my real purpose here is not to praise interaction but to bury it—and, ultimately, to go back and revise some of these motivations. So it's time to move past what seems right about interactionism and get on to what is actually wrong with it.

3.4 INTRODUCING THE CASE AGAINST INTERACTIONISM

What's wrong with interactionism? Here is what I claim to be the underlying problem, in broad strokes: interactionism tells you to *both* correctly assess the first order evidence *and* then go on to take account of the higher order evidence, possibly changing your doxastic attitude as a result. So, for instance, the weighing view tells you: see where the weight of the first order evidence falls, and then move to the (possibly different) confidence obtained by summing in the additional weight of the higher order evidence. But why ever change one's view this way? If the first order evidence is available for responding to in the first instance, then that is straightaway what one should do and nothing further. Given that interactionists take the first order evidence to matter, they must take us to be capable of responding to it; but once they take us to be capable of responding to it, there is no intelligible rationale for why we should ever go on to do anything else. So interactionist ask us to combine two factors when, properly conceived, one simply ought to trump the other. But this advice doesn't make any sense.[42]

That, anyway, is the critique I am going to be defending. How will I do so? My strategy begins by demonstrating what I take to be a manifestation of the above problem. I show the existence of cases where one cannot reflectively follow the interactionists' advice: following it requires thinking that one isn't. The reason is that if you *did* understand that you were following it, you would promptly stop and do something else instead. So interactionism sometimes requires epistemic akrasia.

It is important to be clear that this demonstration is not supposed to refute interactionism on its own. It is controversial whether, and when, akrasia may sometimes be epistemically required. And some interactionists are willing embrace the akratic implications of their view (as I will discuss). Rather, the point will be to show that *this* akrasia is unattractive, that it is arising as

[42] A critique like this is mentioned in passing by White in his (2009), where he wonders in a footnote why it should be that the first order evidence is allowed to do *some* work, but yet not all of it? Similarly, Schoenfield in her (2014) also briefly raises a related worry for a particular brand of interactionism —"e-calibrationism"—when she asks why, within this theory, a number representing *the first order evidence* should be modified in light of another number representing *the agent's reliability at judging the first order evidence* (which, on that brand of interactionism, is what the higher order evidence amounts to). The second thing isn't *about* the first thing, so why should it modify it? I take both White and Schoenfield to be correct: if the first order evidence is allowed to do *any* work it must be allowed to do it *all*, and it is for this reason that the interactionist has trouble specifying interpretations of their variables on which the modifications they propose make sense. I see myself in this paper as trying to develop a spiritually kindred critique both as fully as possible and at the highest level of generality possible.

a symptom of the underlying more serious problem that interactionist norms don't make sense. And so I will also argue that too, once I have finished establishing the phenomenon.

3.5 AKRASIA IN SUCCESS ENTAILING CASES

We can engineer conditions under which the interactionist will be afflicted with akrasia. Here is my recipe. Start a thought experiment with an initial use of *success entailing reasoning.* Follow it with subsequent *overwhelming defeat.* Then stipulate that this happens under conditions of *reflective application.* The result will be that actually following interactionism requires believing that you haven't. So you will wind up being required to hold a doxastic state you take to be irrational.

Ok: what do I mean by 'overwhelming defeat,' 'success-entailing reasoning,' and 'reflective application?' I have already mentioned the first two in passing, but here they are again. An 'overwhelming defeat' case is one where the higher order evidence is so overwhelming that, at the end of the day, one ought to have a very low confidence that one's initial answer was correct (even if one's answer is in fact correct). And a 'success-entailing reasoning' case is one where correctly assessing the first order evidence entails getting the right answer.

To fix things in our mind, we may recall that the dining table case has these features. The logical reasoning used ('because the central connective used is…' 'if the antecedent is true then…' etc.) is such that one cannot get an incorrect answer by reasoning correctly. So it is a case of success entailing reasoning. And the facts that one was drunk and that one's companions unanimously came to a contrary result lead to overwhelming defeat—in the face of all that, one should become almost certain that one's initial judgment was wrong (if one disagrees that this is sufficiently overwhelming, feel free to keep larding the case with more such hostile higher order evidence until one feels inclined to agree: add more friends, more compromising substances, etc...). So the dining table case is (or easily can be) one of overwhelming defeat.

What of the third condition—what does it mean to say that it is also a case of reflective application? A case of ‘reflective application is one where the subject is epistemologically sophisticated enough to know the true epistemic theory plus relevant background facts from pure epistemology. Additionally, they have adequate introspective capacities to know, at least in broad strokes, how they are reasoning through the case at hand. So they know both (as epistemology) what is required and (as introspective psychology) the general features of how they are trying to discharge that requirement.

We did not include this in our initial description, but on natural fillings in the dining table case also satisfies this condition. We may suppose the protagonist is an excellent epistemologist; we may also suppose that they are an ordinary human adults, with ordinary introspective access to their thought processes. So, they know the true epistemic theory, and as they work their way through the case they (mostly) know what it is that they’re doing.

So, the dining table case has those three features: *success entailing reasoning, overwhelming defeat,* and *reflective application*. Now suppose that interactionism is true, and that I, flawless agent that I am, have made my way through the dining table case by perfectly following it. First, I correctly assessed the first order evidence. So I concluded that the formula was a tautology. But then, when I was confronted with overwhelming defeat—my friends, the wine, whatever else—I properly lost almost all my confidence, again just as I should. So, at the end of the day I wound up with a very low confidence that the formula is a tautology. Now ask: what happens when I reflect on what I have done? I say: I must become almost certain that my ultimate doxastic state is irrational.

How do we derive this result? We do so by way of the following overlapping material conditionals, each of which I am in a position to know in the case as described (I will shortly return to explain *why* I am in a position to know each, but for now we are just getting the structure onto the table): 1) My current attitude is rational → 2) In forming my current attitude, I have employed an accurate assessment of the first order evidence → 3) The first order evidence supports believing the formula is a tautology → 4) the formula is a tautology.

We may suppose that since this is a case of reflective application, not only am I in a position to know these conditionals, but that I have actually thought about and *do* know them. If this is so, then maintaining probabilistic consistency forbids me from investing more confidence in any of their antecedents than their consequents. Since the conditionals overlap, this constraint applies to the ultimate antecedents and consequents as well. So, probabilistic consistency forbids me from being more confident in 1 than in 4. But, since this is a case of overwhelming defeat, my confidence in 4 has to be very low. And since my confidence in 1 cannot exceed my confidence in 4, it follows that my confidence in 1 has to be very low as well. And that was the result we were looking for: my confidence in the rationality of my current attitude has to be very low.

Most of the action here rests on the claim that I am in a position to know these conditionals. So, as promised, I now supply the rationale for why that is so.

Start with 1 → 2, which states that in order for my current attitude to be rational it must be formed on the basis of an accurate assessment of the first order evidence. This is close to, although not quite, just a restatement of the interactionist claim that the first order evidence matters. After all, if the true epistemic theory takes an input, and you use the wrong input, you get the wrong result—right? Well, not quite. Perhaps, e.g., you have managed to make some other complimentary errors elsewhere, and the mutual effect of these errors is to cancel out and thereby nonetheless return the answer you would have gotten by doing everything right. Possible, but the default assumption is that this isn't so, and similarly the default assumption is similarly that the use of a misrepresentation in attempting to comply with interactionism will subsequently compromise the result. And since these considerations, plus the truth of interactionism itself, are all background facts from pure epistemology, it follows that as I can know them all just by way of the epistemic competence specified in *reflective application.* So I can know 1 → 2 just by virtue of that competence.

Now consider 2 → 3, which states that in order for the assessment of the first order evidence I have employed to be correct, it must be that the first order evidence supports believing the formula is a tautology. Why am I in a position to know this? Well, as per the description of the case I have in fact flawlessly complied with interactionism. And this means that I must be basing my current confidence in a correct assessment of the first order evidence. In this case, that is an assessment which holds that the first order evidence supports the formula being a tautology. Given that I possess the introspective capabilities specified in *reflective application*, I am in a position to know that this is the assessment I am employing. So I can also know what follows immediately from my employing that assessment, namely that my assessment is accurate only if the first order evidence is as it represents.

Finally, consider 3 → 4, which states that in order for the first order evidence to support the formula being a tautology, the formula must be a tautology. This follows from the use of success entailing reasoning: for there to be a first order route to the conclusion being a tautology it must be a tautology. Since this feature of logical reasoning is another fact from pure epistemology, I am again in a position to know it on the basis of the background epistemic knowledge specified in *reflective application.*

So that, then, is the explanation of how I can know all three conditionals. And once they are in all place their upshot is that my confidence in 4 forms an upper bound on my confidence in 1. My confidence in the truth of my original answer, after discovering the overwhelming defeat, also forms an upper bound on my confidence in the rationality of my new answer. Since overwhelming defeat can force my confidence in my original answer arbitrarily low, my confidence in the rationality of my new answer can be correspondingly forced arbitrarily low. And that, again, is our akratic result.

I think it's clear what's going on here: in the case at hand interactionism requires me to first accurately assess the first order justification and then go on to abandon the resulting doxastic state in the face of overwhelming defeat. But since I began by using success entailing reasoning, there is no daylight between that initial judgment being first order justified and it being true. So in order to take new doxastic state I obtain post-defeat as *true,* I have to also take it to have been what was supported by the first order evidence all along.

But even as I change my view on where the first order evidence has always lain, I cannot correspondingly switch the assessment of it that I employ in whatever reasoning processes brings me into compliance with interactionism—I am still, as before, required to keep on using the same (correct) assessment I began with, because interactionism requires I take account of the first order evidence *itself* and not merely my (possibly wrong) view of it. So the change in my view of the first order evidence does not induce a corresponding change in my use of the first order evidence. So a gap opens up between the two. I am required to simultaneously believe that the first order evidence supports one thing but yet treat it like it supports another. And that disconnect is what compels me to conclude that the (actually rational, flawless) doxastic state I have adopted is irrational.

3.6 OBJECTIONS AND REPLIES

Here I canvass some assorted ways an interactionist could try to block the foregoing analysis of the dining table case, arranged from least to most promising.

I do not think there is any promising way in which the interactionist can contest the existence of cases where we use success entailing reasoning, or of cases where we are faced with overwhelming defeat, or the fact that they sometimes go together. We simply do use success entailing reasoning, sometimes run into overwhelming defeat, and sometimes do both at the same time. Indeed, this was supposed to be a key motivation for interactionism. It, unlike right reason, could accommodate the obvious fact of overwhelming defeat, complete with bog standard example that itself occurs within a success entailing context, e.g. ‘a million mathematicians advise you that you were wrong about…’ So objecting to either of those features seems like a no go.

Rather, if one of the features of the case can be legitimately objected to it seems that it is *reflective application*, which states that I have 1) relevant background knowledge from pure epistemology and 2) introspective knowledge of the gross features of my own reasoning process. Now, the first clause again strikes me as difficult to reject. After all, why shouldn’t I be able to have background knowledge from pure epistemology? It would be quite *ad hoc* if interactionists had to insist that I spontaneously forget the lessons of the seminar room—anodyne epistemic facts on the order of ‘logical reasoning is success entailing’—in order to pave the way for compliance with their preferred theory. Rather, it is the second clause that is more vulnerable, the clause that specifies that I have introspective knowledge of my reasoning processes.

Now why was that introspective knowledge important? It was required in order to let us conclude that if I was in fact employing some particular assessment of the first order evidence in my efforts to comply with my interactionist view, then I knew that was the assessment I was employing. So, given that I *was* feeding "the first order evidence supports believing the formula is a tautology" into the interactionist reasoning process (a necessary condition on doing it right), I also knew that's what I was doing.

Now when we think of explicit, conscious reasoning, this sort of self-awareness is pretty trivial. If I think out loud to myself "the forensic evidence points toward Bob, but the testimony points towards Sherry, the forensic evidence is more important, so Bob probably did it" then I am fully aware of whether I am taking the forensic evidence to point toward Bob or Sherry. But not all reasoning need be this explicit. And perhaps the interactionist might defend their view by claiming that the reasoning in question is neither explicit nor introspectively available. Now, we would again want some principled reason why this reasoning not only *can* but *must* be carried out subpersonally—again, we would not want it to be that a person loses introspective capacities just for the convenience of interactionist accounts. But here it is at least more plausible that there could be such a principled reason. At the very least, the subject of this introspective knowledge (what a person is doing in their reasoning) is facially linked to the cloudy phenomenon in question (controlling for unreliability in reasoning), so there may be some promising rationale to draw here.

In response, though, I would emphasize the following two things. The first is that introspective knowledge required to get akrasia going here is quite spare. Even when we do not know *entirely* how we are responding to evidence at a subpersonal level, we do often know the gross features of our reasoning process. Upon reading several editorials about the state of the election I may come to a firm overall opinion on which candidate will win without knowing *exactly* how I in fact weighed and compared each article. But nonetheless, even if I don't know exactly how much or how I am taking each article to support each candidate, I will typically at least know which candidate I am taking each article to support. That is to say, I will be able to answer questions like "if this was the only article you read, would you think Trump was going to win?" and "if you hadn't read this article, would you be more or less confident Trump was going to win?" But even this very unspecific access to my own reasoning is all I need in order to wind up in the pickle described in section 3.5. All I need to know is that my reasoning process is employing an assessment of the first order evidence on which it supports the claim that the formula is a tautology *to any degree at all, as opposed to the opposite* in order to know that my assessment of the first order evidence can be accurate only if the formula is in fact a tautology. That is a gross, easily accessible feature of my reasoning process, not a finicky, obscure one.

Furthermore, the limited introspective capacities required for the above argument against interactionism *in general* are actually significantly more robust than those required for the argument against any particular interactionist view. Once we specify the particular interactionist view, its contents become yet another background fact in pure epistemology for the agent to know. And, upon knowing them, the agent can come to similar conclusions with much less self-knowledge. For instance, on the accounts of *all* of (Kelly 2010), (Sliwa and Horowitz 2015) and (Christensen forthcoming) I will often be able to infer just from the combination of an uncontentious description of the higher order evidence and the final verdict I arrived at to the unique assessment of the first order evidence that I must be employing if I am to be following the view.[43] Once we actually start characterizing the processes involved in these interactionist views I can proceed by inference from more remote indicators to what I *must* have been doing if I am to have conformed to the view as so specified. And those remote indicators, like the mere knowledge of one's final confidence in one's answer, can be exceptionally weak. So I don't find the *reflective application* condition a very promising place for the interactionist to push back either.

The final objection I will consider does not try to contest the akratic verdict in the dining table case. Rather, it tries to contain the damage. It concedes that success entailing reasoning may cause problems, but counters with the dual assertions that such reasoning is relatively rare and anyway independently puzzling. After all, how bad is it to be flummoxed by the special features of the epistemology of logic? They might be in good company there. In order to forestall this response, I now move on to generalize the case. Starting with success entailing reasoning makes the issue especially clear and is a helpful way to introduce the problem, but it is ultimately dispensable: the combination of *overwhelming defeat* and *reflective application* is already enough to force an akratic result even when the reasoning in question isn't success entailing.

[43] For Kelly, the higher order evidence will often be balanced in a straightforward way: for instance, when I judged P and my peer judged ~P that will present symmetrically balanced higher order evidence. If, in the face of that symmetric higher order evidence, I ultimately become more confident in P than ~P, that can only be on the basis of having employed an assessment of the first order evidence as having supported P rather than ~P. For Horowitz and Sliwa and Christensen, on the other hand, there will be cases where the estimate of one's reliability in a given situation is very straightforward. And when one is in such a case and goes on to adopt some doxastic state, one will (at least often) be able to reverse engineer the assessment of the first order evidence such that it would, when calibrated by that reliability estimate one knows to be appropriate, result in the doxastic state one has just adopted. For instance, for Horowitz and Sliwa, if I am in a situation where my expected reliability is .6, and I am .6 confident of P as against ~P, then I must have assessed the first order evidence as supporting P if I am to have followed the view.

3.7 AKRASIA IN GENERAL

Imagine things proceed as in the dining table case, but with the following substitution: instead of wondering whether some formula is a tautology, I wonder whether some e-mail I have just received indicates that my coworker is angry with me. So I begin by initially (and correctly) assessing that it does on first order grounds—on the basis of the word choice, the decision to raise and omit certain details, and so on. However, I then come into possession of powerful higher order evidence that overwhelmingly defeats that initial assessment. Perhaps all my dining companions rebuff me, unanimously agreeing that I have always an inexplicable grudge against this person, I am mean when I've been drinking, and so on, again with whatever here is enough in our sensibilities to really be overwhelming. So abandon my initial assessment, and adopt a very low confidence that my coworker is angry with me.[44]

In this situation, we can still write out the same conditionals. The difference will be that the last one fails to hold. 1) My current attitude is rational → 2) In forming my current attitude, I have employed an accurate assessment of the first order evidence → 3) The first order evidence supports believing my coworker is angry —/→ 4) My coworker is angry.

Because reasoning from the contents of an e-mail is not success entailing, it is no longer true that the first order evidence supports believing my coworker is angry only if they are angry. And given that the conditional 3 → 4 fails to hold, so too is my confidence in 4 no longer an upper bound on my confidence in 1. So I can get the desired combination of respecting overwhelming

[44] We may clean up some potential complications by stipulating that A) my coworkers do not actually look at the e-mail in order to make this judgment, so their reactions are not functioning as further first-order evidence with respect to what its contents indicate—they are *just* telling me something about how unreliable they know me to be at this type of thing, and B) the base expectation that my coworker is angry with me at any given time, absent evidence one way or the other, is very low, and so very low confidence is the state I am returned to once I entirely stop investing any credit in my dubious initial assessment of the e-mail.

defeat by being almost certain my coworker isn't angry, and yet still maintaining a healthy confidence in the rationality of my doxastic attitudes.

Upon closer examination, though, although such an attitude would be consistent, it would not be desirable. Note that the conditionals 1 → 2 and 2 → 3 still hold, since nothing in their justifications turns on the use of success entailing reasoning. But given that they still hold, that means that my confidence in 3 is still, even now, an upper bound on my confidence in 1. So the only way that I can be highly confident in 1 is by being equally highly confident in 3, and correspondingly much more confident in 3 than in 4. That is to say, the possibility that opens up when I stop using success entailing reasoning is that *although* the first order evidence supports believing my coworker is angry, they actually aren't. It is only by becoming highly confident that this is the case that I can respect defeat while avoiding akrasia

The problem is that this response makes no sense in the case as described. My companions' testimony, the wine, the overwhelming defeat generally—these are not giving me evidence that my original evidence *is misleading.* They bear not at all on whether a correct interpretation of this e-mail would yield a true verdict. Rather, they are giving me evidence that *I failed to correctly evaluate it* (whatever it might support, and regardless of whether that support is veridical). If I responded to overwhelming defeat in the manner depicted, it would represent taking the testimony of my companions to be testimony to the effect that I had been faultlessly misled by my evidence. But, again, that would be a facially crazy response to the evidence being presented.

Something similar can be said with respect to an issue we shelved earlier. I have been taking it to be the case that getting an input wrong, i.e. employing an incorrect assessment of the first order evidence, correspondingly means getting the output wrong, i.e. coming to the wrong final confidence. In doing so, I have been bracketing the possibility that, through some fortune, multiple mistakes all cancel each other out and lead to getting the output right even though an input was wrong. This is at least conceptually possible, and so far all I have had to say against it is that the default assumption is that it isn't so. Nonetheless, the appeal to defaults may seem hand-wavy in this context. Fortunately, we are now in a position to say something more significant about why we may legitimately bracket this error.

What we say is very similar to what we have just said about the possibility of misleading evidence: although it is possible to use a mistaken input and yet still get the right answer anyway, this is not what I am getting evidence of when my friends tell me that I am wildly unreliable and I realize that I've had too much to drink. I am getting evidence that I've made a mistake, but nothing about it even hints that this was a mistake that fortuitously managed to produce the right result. So although becoming confident that this is what happened would allow me to maintain a comfortably high confidence in 1 consistently with my very low confidence in 4, it would do so only by treating my higher order evidence as indicating something it simply doesn't.

So I take it that these options are non-starters. The important upshot, then, is that the tension between interactionism and *reflective application* is really quite direct. It can be avoided only by way of doing great violence to the intended interpretation of the case. The interactionist's problem with akrasia is not a curiosity, but rather directly follows from what it is that higher order evidence was supposed to be about in the first place.

3.8 HARMLESS AKRASIA

So, interactionists get lead into akratic results sometimes. What is the significance of this result? It is worth noting, preliminarily, that some authors reject 'level connections' entirely; that is to say that they allow what you are rationally required to believe to swing entirely free from what you are rationally required to believe you are rationally required to believe. Weatherson, for instance, holds that these are just different propositions and your body of evidence can simply diverge with respect to what it says about each (Weatherson ms.). On a view like this, akrasia is not even *prima facie* an indication of something gone wrong. But this is not, on the face of it, a happy position for an interactionist to take. Interactionists are committed to the higher order evidence mattering. But it is hard to see how higher order evidence could matter if not by way of *some* traffic between one's beliefs about what is rational and one's beliefs about first order matters.[45]

And indeed, this hard line has not been the favored response of those interactionists who have noticed the akratic consequences of their views. Rather, they have taken an eminently more sensible tack. They start by identifying something bad about akrasia. They then show that although their view licenses some akratic-looking combinations, none of them have that bad-making property. So although their view does sometimes require akrasia, at least it's not the bad kind.

So, for Sliwa and Horowitz in their (2015), the problem with akrasia is that it sometimes licenses a form of intuitively objectionable 'bootstrapping' reasoning—reasoning wherein a person comes to increase their confidence in the reliability of their faculties just on the basis of repeatedly deploying those selfsame faculties with no independent checks.[46] And Christensen in his (forthcoming) diagnoses akrasia as impermissible when it indicates inaccuracy, *prima facie* permissible when it doesn't, and he further suggests as promising the use of a particular principle

[45] Certainly, (Kelly 2010)'s claim of a downward push coming from epistemic propositions could not survive a complete divorce of the 'levels.'

[46] 'Bootstrapping' reasoning has been discussed exhaustively in the context of objections to reliabilist theories. See, e.g., (Cohen 2010a).

for establishing level connections, namely, the New Rational Reflection principle originally formulated by Elga in his (2013).[47]

Having done that groundwork, these authors are then in a position to draw favorable contrasts for their view. So, Horowitz and Sliwa point out that their particular brand of interactionism is guaranteed to never permit bootstrapping reasoning. So the bad-making property is avoided. And Christensen notes that the cases where his view leads to the sharpest akrasia are ones where rationality and accuracy predictably diverge, and so believing yourself irrational need not require believing yourself inaccurate; in the less sharply akratic cases, he notes that his view appears to avoid violations of New Rational Reflection. Again, the bad-making properties are avoided. We are left with only harmless akrasia.

These responses follow an eminently sensible defensive strategy. And I find them plausible so far as they go, in terms of defusing specific threats: indeed these views do not license bootstrapping, nor do they appear to violate New Rational Reflection. Rather, my counterpoint will be in some ways orthogonal. As I said earlier, I believe the akrasia being engendered here is a symptom of a deeper problem. If I am right about the problem, then it is untouched by situational defenses of some forms of akrasia. Showing that is the task I turn to now.

[47] That principle states that my actual credences should satisfy the following equality, for all propositions A and all probability functions PR: Cr(A| PR is ideal) = PR(A | PR is ideal)). Colloquially, it expresses an attitude of deference to opinions proportional to the extent to which you are confident that a perfectly rational advisor *who knew that they were perfectly rational* would hold those opinions—the latter clause is what sets it apart from ‘old’ rational reflection, an emendation which among other things enables it to handle some purported counterexamples.

3.9 THE PREEMINENCE OF FIRST ORDER EVIDENCE, WHEN CONSIDERED AS SUCH

We can begin by observing a curious feature of the relationship between first order and the higher order evidence. Return to the dining table case, where by description my initial assessment of the first-order evidence is correct; I have adopted the doxastic attitude it supports. After forming that assessment, I go on to take account of the higher order evidence too. Question: can the act of doing so *improve* my doxastic attitude? Answer: it cannot. On the other hand, can it make it worse? It can. So the very best that taking account of the higher order evidence can do is fail to make things worse.

In the first variation of the dining table case, the holding of this pattern is just a trivial consequence of the fact that I've used success entailing reasoning. After all, correctly having reasoned to the formula being a tautology means that, indeed, it is a tautology. So, having correctly responded to the first order evidence already leaves me with the best possible attitude. And this then means that any further consideration, wherein I go on to respond to the higher order evidence, can do only one of two things: nothing, because it does not change my attitude, or something, because it changes my attitude for the worse. So after having responded to the first order evidence correctly, further consideration of the higher order evidence has no upside.

But this also holds in the second variation of the dining table case, where we move on to consider contexts where our reasoning is not success entailing. Now, it is no longer *guaranteed* that moving from the attitude supported by the first order evidence will be all downside and no possible upside—after all, the first order evidence may have been misleading (since no longer success entailing) and so it may have supported the wrong thing. But even though this is possible, the presumption should still be that moving away from the attitude it supports is a bad deal. We could argue for this on the basis that we should presume that the first order evidence isn't misleading, because there is some general rational expectation that evidence isn't misleading. But I think there is a more direct route. To see it, we just have to remind ourselves of what higher order evidence is supposed to be about.

Higher order evidence, as we have characterized it, is evidence that affects our confidence in our final answers not directly, but rather *only* by way of intermediately affecting our confidence in the correctness of our first-order reasoning. The fact that we were inebriated, that our friends came to different verdicts, that we have some practice at these sorts of reasoning tasks, and so on and so forth—these facts matter, if they do, because they bear on whether we were successful in forming our initial judgment.

But given that this is what higher order evidence is about, that again leaves us with a dilemma. If you have correctly assessed the first order evidence, then the higher order evidence can only go on to do one of two things. It can (veridically) point to the conclusion that you correctly assessed the first order evidence, in which case it doesn't call for revisions to your attitude. Or, it can (misleadingly) point to the conclusion that you didn't correctly assess your first order evidence, in which case it may call for revisions. But then it follows that every case in which revisions are called for is one in which those revisions are based in misleading evidence. Revisions based in misleading evidence should, as a general matter, be presumed to worsen our doxastic attitudes. So even in cases which were are not success entailing, it will still be true that if one has correctly responded to the first order evidence, then further adjustments in light of the higher order evidence won't make things better, but may make them worse.

The point here is that higher order evidence matters, if it matters, only because it is a guide to what the first order evidence actually supports. But it is quite generally true that once you specify the actual facts about an object of interest, there is nothing to be gained from further evidence whose upshot is entirely confined to its role as a guide to those already specified facts. So what we see when we examine the relationship between the first order and higher order evidence is just a specific instantiation of that general fact; once the first order evidence is gotten right, higher order evidence has nothing to offer.

3.10 AN UNFLATTERING PARALLEL

If I am right about the preeminence of first order evidence with respect to higher order evidence, and if I am right about why it is preeminent, then there is serious trouble for the interaction. For it is quite generally true that there is no sensible perspective from which we can take both *a thing, considered as such* and *a guide to that thing* as meaningful inputs to our decision making. Now, that's pretty abstract and programmatic, but I believe it can be clarified and borne out by way of example.

Suppose that we are interested in whether some patient has a lymphatic disturbance. If the patient has swollen lymph nodes, then that entails that they have a lymphatic disturbance (since swelling is one way for lymph nodes to be disturbed). Furthermore, it is also true that sweaty palms are evidence that a person's lymph nodes are swollen. So, sweaty palms are the sort of thing we can use as a helpful guide to lymphatic swelling in those conditions where it is not possible to measure directly. However, sweaty palms are interesting *only insofar as* they play that evidential role with respect to whether there is any swelling; they have no other relation to lymphatic disturbance. If we'd like we can use probabilistic terminology here, and we can say that the swelling of a patient's lymph nodes *screens off* the sweatiness of her palms with respect to whether they are suffering a lymphatic disturbance.[48]

Now consider the "confused doctor" theory. On the confused doctor theory, one's diagnosis of a lymphatic disturbance ought to be the product of two factors: the swelling of their lymph nodes and the sweatiness of their palms. First, one should check the swelling of their lymph nodes and set one's confidence accordingly. Then, one should adjust that confidence up or down depending on the sweatiness of their palms. In particular, the theory holds that the complete absence of sweat triggers an analogue of overwhelming defeat: at a certain level of dryness, one must abandon almost all confidence that there is a lymphatic swelling.

Now this is, on its face, clearly not a good theory—after all, we have already specified that there is nothing to learn from sweatiness once the swelling is accounted for. And given that this is so, adjusting on the basis of sweatiness after having already checked for swelling is senseless. But let's pause to see what would happen if a doctor *were* actually to go through a diagnostic procedure while employing this theory. So, suppose that, as a doctor, I see patient P. First, I feel P's lymph nodes and notice that they are quite swollen. So I initially adopt a high degree of confidence that P has a lymphatic disturbance. After all, I feel that the swelling and swelling *just is* a type of disturbance. But then I feel P's palms. P's palms are totally dry. So I abandon almost all of my confidence that P has the disease and end up with a low ultimate confidence that P's lymphatic system is disturbed.

[48] I take the notion of screening from (Weatherson ms), which attributes it to Reichenbach. In general, A screens B with respect to C when P(C|B) > P(C) but P(C|A&B) = P(C|A). Weatherson also uses this notion to discuss the relation between first and higher order evidence and I have found his discussion helpful. As I will reveal, I think he is in one way substantially right, though wrong in another.

Now, if we take all of this as fixed, what happens? I am faced with conditionals substantially similar to those I had before: 1) My current attitude is rational → 2) In forming my current attitude, I have employed an accurate assessment of the patient's lymphatic swelling → 3) The patient's lymph nodes are swollen → 4) The patient has a lymphatic disturbance

The confused doctor theory requires me to end with low confidence in the patient having a lymphatic disturbance, and so given that I have flawlessly followed it that is what I do. But I also know that if the patient's lymph nodes were swollen then that entails that they have a lymphatic disturbance. And I know, via appropriate introspection, that I did indeed treat them as having swollen lymph nodes; I remember how I checked their swelling first, set my confidence high, and then only lowered it once I went on to discover their sweaty palms. So I know that I have successfully followed my theory only if their lymph nodes really were swollen. But that means that I have successfully followed it only if the patient actually has a lymphatic disturbance, which is the very result that the theory requires that I must be confident doesn't obtain. So, to follow the theory I must be confident that I haven't.

What we are seeing is that these propositions have the same structural interrelations as those propositions that we had in the first version of the dining table case. Just as the first order evidence there entailed the formula was a tautology, here the existence of lymphatic swelling entails the presence of a lymphatic disturbance. When we impose over this relationship a theory which requires us to get that entailing evidence right, but then go on to adopt some other doxastic state than the one it supports, we get resulting akrasia.

The case also proceeds in parallel when we alter it such that the connection between 3 and 4 is no longer deductive. So, instead of being interested in whether the patient has a lymphatic disturbance (which is entailed by lymphatic swelling) suppose that we are interested in whether the patient has a rare tropical disease (for which lymphatic swelling is merely evidence). Now, so long as we specify that, on our intended interpretation, that the sweatiness of the patient’s palms are evidence for this tropical disease *only* insofar as they are evidence for lymphatic swelling—just as higher order evidence is interesting only insofar as it is a guide to the support of the first order evidence—we again obtain the same structure. There will be two ways that one can follow the confused doctor theory, and hence end up with a very low confidence that the patient has the tropical disease: either one can become very confident that one failed to follow the confused doctor theory, or one can become very confident that one is in a case where the lymphatic swelling is present but is misleading with respect to the presence of the tropical disease. Opting for the latter option would involve treating the sweatiness of the palms as evidence either *about whether the lymphatic swelling is misleading* or more directly about *whether they have a tropical fever*, but these are nonstarters on the interpretation of the sweatiness we stipulated. So again, we are instead forced into the first option. To follow confused doctor theory requires believing that one isn’t following it.

What is the point of this parallel? What we are seeing here is that the akrasia to which interactionists are committed doesn’t come from the *topic area* of the claims being adjudicated. One might have thought, rather naturally, that interactionists are forced into akrasia because they are trying to handle claims that are *about* their own rational performances. But this isn’t so. We can replace epistemic terms ‘the support of the first order evidence’ and ‘higher order evidence’ with non-epistemic terms like ‘swollen lymph nodes’ and ‘sweaty palms’ and structurally analogous theories for handling them will still get the same interactionist-style result. So this isn’t an unfortunate but difficult to escape consequence of self-reflective epistemic theorizing. Instead, it's the general consequence of theories which try to simultaneously attend to two inputs when one of those inputs is properly viewed as *only* relevant as a guide to the other.

Indeed, we can notice that the akrasia generated by the confused doctor theory doesn't carry either of those distinctive flaws noted earlier. It does not allow for the possibility of bootstrapping into undeserved confidence in your diagnostic abilities.[49] Nor does it introduce any violations of New Rational Reflection.[50] But for all that, I doubt very much that anyone would be much inclined to defend the confused doctor theory. Rather, the theory wears its flaw right on its sleeve. It just doesn't make any sense to revise your diagnosis on the basis of the sweaty palms after having already checked for the lymphatic swelling. The result, then, is that you're forced into akrasia. Since what your theory is demanding you do doesn't make sense, you have to come up with some interpretation of your actions on which you aren't actually doing it. But then, since it is also what you take to be required, that means that your own interpretation of your actions will be one on which you aren't doing what's required.

This is what I meant earlier when I claimed that the problem with interactionism is not that it leads to akrasia *per se.* I take it that the problem with interactionism is that it doesn't make sense. Akrasia is just the most visible manifestation of that problem.

[49] Nothing in the story as told obviously implies anything particular about the confused doctor's beliefs about their diagnostic reliability. For all we've said, they could both be and stay maximally confident that their diagnoses are always correct (not necessarily rational, but correct in the end), and hence have no room to bootstrap.

[50] How could confused doctor theory, but not interactionism, require such violations? The structural relations between the theory and the propositions are the same, and none of the contents of the propositions that were changed are about the rationality of any particular credence. If there was no failure of new rational reflection before, it's unclear how the change could force one.

3.11 TOWARD A RESOLUTION

If we think of me, qua doctor, as having direct access to the swelling of the patient's lymph nodes, then I should simply check them and pay no attention to the sweatiness (or lack thereof) of their palms. Perhaps I just reach out and feel: yep, swollen. Now having done that, why would I even bother to check for sweaty palms? I just felt the lymphs and they're swollen. However, there is a way that we can get ourselves into the state of mind where checking the palms tells us something—and that's by imagining that we *don't* have direct access to the swelling of their lymph nodes. Reinterpret what I respond to as a good, yet fallible indicator, as would be, say, my memory of having checked, that the neck *felt like* there was swelling, or the listing of swelling on a patient's chart that I've just been handed. All of a sudden, now we can make sense of learning something from the sweaty palms in a way that we couldn't before. That's because now there's room for doubt about the swelling, doubt that the facts about sweaty palms can productively interact with. Dry palms may, then, make me ever so slightly less confident that this good yet fallible indicator has really told the truth in this case. And I lead with this because I think that the same holds of the relation between first and higher order evidence.

Recall, way back when we were motivating interactionism, how we went through some arguments *against* taking account of each type of evidence. First, we argued against taking account of higher order evidence on the grounds that it wasn't about the right sort of thing; first order evidence fully settles the question of whether a formula is a tautology and higher order evidence has nothing further to offer *about whether that formula is a tautology*. Second we argued against taking first order evidence into account on the grounds that it wasn't accessible in the right way: the point of these problem cases is that I am unsure of what the first order evidence supports, and so to productively manage that uncertainty I need to deploy tools which do not themselves presuppose that support to lie in any particular place. Now, I think that what we are seeing here is that both these arguments are correct. If we take up a perspective which includes the first order evidence itself in the description of the situation we face, then there is nothing to do but simply adopt the doxastic state it supports and higher order evidence has no further role to play. But such a perspective is, for obvious reasons, utterly lousy for managing our uncertainty across what the first order evidence supports. On the other hand, we can take up a perspective useful for that task, one informed by the higher order evidence. But if we do, it can at most contain indicators for what the first order evidence supports—initial judgments, intellectual appearances, and so on—and not the first order evidence itself.

Okay: we can talk about different perspectives. But why would we, and what ice does this cut? Now, I am not going to offer anything like a full theory here, but I will at least wave my hands and say something suggestive. Namely, I want to suggest that separating out these different perspectives can help us do important meta-epistemological work.[51] Each perspective has important uses, and this shows us that we need multiple types of normative concepts. I suspect that there are a cluster of largely third-personal, evaluative concepts that benefit from being conceptualized by way of the first order perspective: they can do important work, e.g., in explaining success, especially relative success, in justifying deference to experts, and so on. But they cannot tell the whole story. We also have a cluster of first-personal, prescriptive concepts that benefit from being conceptualized by way of the higher order perspective: they can do important

[51] Schoenfield has done excellent work in actually developing out an account of the first order and higher order perspective (Schoenfield forthcominga). She does so in terms of plans *that are good to follow* as against plans *that are good to try to follow*. I find myself mostly in violent agreement with her. See also her (2012) for an earlier exploration of similar themes, there fitted to some puzzles surrounding precision as a rational requirement for credences.

work, e.g., in explaining how an agent may form 'epistemic plans,'[52] and in rendering rationally intelligible the performances of others.

Furthermore, once we have the ability to appeal to multiple notions, then that also holds out promise for making headway on dissolving the argument that initially pushed us toward interactionism in the first place. We may, for instance, be able to explain overwhelming defeat in terms of the higher order perspective and its attendant concepts. At the same time, we may also be able to explain what is wrong when a person flubs a judgment against a background of competence using the first order perspective and its attendant concepts. And we needn't be forced to find a way to make these perspectives and their concepts agree. We might, for instance, say that a person who flubs their judgment against a background of competence thereby arrives at a *rational* belief that is nonetheless *unsupported by their evidence*; or, correspondingly, that a person facing overwhelming defeat has an *irrational* belief despite it being fully supported by their evidence.

It is not my purpose, though, to work out such a multi-concept story in this small remaining space. Rather, I count this paper successful merely if it motivates us to seriously pursue such an account. Because, if my arguments are correct, then interactionists' attempts handle both phenomena with a *single* concept hold no promise.

[52] See (Schafer 2014) and (Schoenfield forthcomingb) for nice recent work deploying and developing 'planning' based concepts. See also (Gibbard 2003)'s exhaustive work developing a planning-based conception of rationality in pursuit of an expressivistically kosher yet compositionally adequate theory of normative language.

4.0 ANTICIPATING FAILURE AND AVOIDING IT

4.1 INTRODUCTION

When we try to figure out what to believe, having evidence is good and having more evidence is even better. We like the following sequence of events: I get some new evidence, and then, on that basis, I update my beliefs. Call this *The Process:* I hold some beliefs P1, P2, etc. on the basis of some evidence E. I acquire some additional evidence, E*, and then on that basis adjust my previous beliefs so as to arrive at some new P1*, P2*, etc.

I take it to be a basic datum in epistemology that The Process is generally good. It improves our beliefs. That's why we go to such great trouble looking for more evidence, and why, when we are faced with decisions of great consequence, it is especially important that they be made on a maximal evidential basis.

I said above that The Process improves our beliefs. What is it to improve our beliefs? One natural standard is matching; beliefs improve when they match the world better. When we work within the full belief framework this can be cashed out in terms of truth and falsity; we say our full beliefs match the world when they're true. But when we focus on partial beliefs rather than full beliefs, truth can no longer play that role directly—on the usual understanding, a degree of confidence, or credence, cannot strictly speaking be true or false. So instead we speak of the analogue, graded notion of 'accuracy;' we say a partial belief matches the world *better* when it's *more accurate*.[53]

[53] To the extent that the differences among them are relevant, I discuss the particular ways in which accuracy can be quantitatively measured in section 4.9.

The platitudes above about The Process generally improving our beliefs are true on either framework; getting and responding to evidence tends to give us true full beliefs, and accurate partial ones. But since this paper is interested particularly in rational degrees of belief, I will focus on the second version. So: The Process is generally good, where by that I mean that going through the process generally increases the accuracy of one's beliefs.

But even if The Process is generally good, it need not be good in every instance. Sometimes it fails, where by failure I just mean the converse of what I meant by success—namely, that it sometimes issues in beliefs less accurate than those with which it began.

In this paper I begin by looking at two ways The Process can fail. The first: I receive misleading evidence. I faithfully update my beliefs in light of that evidence, but because it is misleading it points me astray. The second: I receive non-misleading evidence. But because I manifest some irrationality, I update my beliefs in ways contrary to the truth that evidence in fact suggests. The evidence points in the right direction, but I go astray anyway.[54]

What should I do if I anticipate that some evidence I'm about to get is such that, in taking account of it, I'm likely to go wrong in one of these two ways—that, one way or another, responding to it by way of The Process will degrade my accuracy? Answer: if I anticipate that taking account of it will lead me astray, I ought to make sure I don't take account of it. Rather, I ought to ensure I avoid or ignore it. To do otherwise would involve taking what I anticipate to be a bad epistemic deal.[55]

[54] This list of ways to fail is not intended to be exhaustive.

[55] The focus on what one anticipates may prime the reader to expect an argument grounded in some reflection principle or other. I explain why that's not the right way to gloss my argument after I've finished giving it, in section 4.15.

In this paper, I show that this seemingly anodyne answer has surprisingly substantial consequences for the current debate on peer disagreement. Specifically, I show how can be used to militate in favor of a 'conciliatory' view. Such views take the discovery of disagreement with one's peers to have serious corrosive effects on one's rational confidence—the version of this doctrine I espouse is roughly Elga's, and the details will be spelled out in section 4.3.

How does the claim that we should avoid taking bad epistemic deals motivate conciliation? It is common among non-conciliatory views to posit some evidence, available in cases of disagreement, that goes unrecognized by the conciliationist and which rationalizes whatever non-conciliatory degrees of confidence they take to be appropriate. But on closer examination it becomes clear that this evidence is such that we can anticipate, in advance, that in responding to it The Process will fail in one of those two ways: it's built into the nature of this evidence that trying to respond to it leads to failure. And if that's right, then, as above, we should not try to respond to it—rather, we should ignore it. Once it is ignored, non-conciliatory views lose their rationale.

The upshot is that conciliatory doctrines about what one ought to believe turn out to be surprisingly autonomous from questions of what the relevant evidence supports.[56] We can allow that the conciliationist is indeed ignoring all sorts of genuine evidential features of the situation, just as their critics allege: we are still left with a surprising and powerful argument that ignoring those genuine evidential features is exactly what they should be doing.

[56] I am indebted to (Schoenfield 2012) for helping me see that there are sensible reasons to multiply our epistemic concepts and separate out what one's evidence supports from what one should believe. See also (van Wietmarschen 2013), which argues in a very different way for a conclusion with a similar ring to it: that conciliation is a bad theory of what the evidence supports, but that it nonetheless may be a good theory of well-grounded belief.

Of course, another option is to insist on the close connection between evidence and what one should believe. That is not my preferred route, but the argument of this paper does not hinge on that choice. If one insists on an inseparable connection between evidence and rational belief, then read the argument of the paper as: non-conciliatory views entail the existence of evidence you should ignore, of which there can be none, and so they must have gone wrong somehow.

Here's the plan: I begin in section 4.2 by discussing some ordinary examples in which we expect The Process to fail, with an eye to motivating the idea that in such cases we ought to avoid basing our beliefs in the problematic evidence. I then move on in section 4.3 to introduce my understanding of what it is to be a peer, as well as the content of the conciliatory doctrine I espouse. Sections 4.4, 4.5, 4.6, and 4.7 introduce alternate non-conciliatory views and explain how they construe the evidence in disagreement; sections 4.8 and 4.9 show how the evidence, so construed, is such that we expect ahead of time that it will cause The Process to fail. With these pieces assembled, I am then in a preliminary position to draw my desired conciliatory conclusion: since we ought to ignore pieces of evidence that we anticipate will cause The Process to fail, and the evidence posited by non-conciliatory views has this feature, it follows that we ought ignore the evidence posited by non-conciliatory views—and, in so doing, embrace conciliation.

Section 4.10 outlines an objection, to the effect that although the evidence posited by non-conciliatory views is such that we expect it to cause the Process to fail *in general*, in each particular instance we are in a position to endorse our responses. Sections 4.11, 4.12, 4.13, and 4.14 answer this objection.

Finally, section 4.15 steps back and situates the argument relative to some more familiar epistemic arguments and claims, before I go on in section 4.16 to conclude.

4.2 ANTICIPATING FAILURE

The Process is generally good. But what precisely do we mean when we say that? We cannot mean that, in each instance in which The Process occurs, the person undergoing it ends up with more accurate beliefs than they had when they started. Sometimes things go wrong, and they instead end up at a worse place than they started. One way this can happen is when misleading evidence leads them astray. Like everyone else, I started this morning believing the world’s nuclear missiles were safely in their silos. But then I saw a report on the local news: they have been launched! So I believed. But in fact the news was being pranked, and there was no launch. In this case, the report was misleading.

Schematically, we may speak of *Failure Due To Misleading Evidence:* I hold some beliefs P1, P2, etc. on the basis of some evidence E. I acquire some additional evidence, E*, and then on that basis adjust my previous beliefs so as to arrive at some new P1*, P2*, etc., which are supported by E*. However, E* is misleading, and P1*, P2*, and etc. are less accurate than P1, P2, and etc. were.

How do we understand the general goodness of The Process such that it’s consistent with its potential to fail due to misleading evidence? It’s normal to want to say something like “misleading evidence is unusual” or “*typically* evidence is not misleading,” or so on. How to cash out these thoughts in a detailed way is not obvious. However, as the interest here is not in a blanket skeptical question, I will take myself to be entitled to the following minimal anti-skeptical conclusion: the rational default is to suppose evidence isn’t misleading. Until one has some reason to think otherwise, one ought to treat one’s evidence as non-misleading, and, concomitantly, one ought to apportion one’s beliefs as it suggests.

This already goes some way toward reconciling the general goodness of The Process with the possibility of failure due to misleading evidence. If the default is to regard evidence as non-misleading (even though sometimes it is), then the default is similarly to regard instances of The Process as improving accuracy (even though sometimes they don’t). And if this is the rational default attitude, we can also rationalize the broad data mentioned in the introduction, namely our general interest in acquiring more evidence and our conviction that acquiring more evidence is especially vital when something important hangs in the balance.

Getting misleading evidence is one way The Process can go wrong. But here is another. Sometimes we acquire perfectly non-misleading evidence, yet we respond to it irrationally. As it were, we acquire evidence that points in the right direction, but we nonetheless go in the wrong one.

For instance: suppose I am superstitious. I begin my day by thinking—work today will be like any other. But on the way to work I notice foreboding clouds on the horizon. I think to myself, "this is a bad sign." I become convinced that I will be fired. Still, the day passes without event.

We can stipulate that the presence of ominous clouds on the horizon was not misleading evidence for the false proposition that I would be fired. This because it was not evidence for the proposition that I would be fired at all; clouds are evidence of rain, and etc., not general heralds of misfortune. It was only by virtue of my irrational superstitions that I came to be convinced on the basis of the perfectly innocent clouds that I would later be jobless.

Again, schematically, consider *Failure Due to Irrationality:* I hold some beliefs P1, P2, etc. on the basis of some evidence E. I acquire some additional evidence, E*, and then on that basis adjust my previous beliefs so as to arrive at some new P1*, P2*, etc. However, despite the additional evidence E* being non-misleading, my response to it is irrational; E* does not actually support P1*, P2*, etc. And, as it so happens, P1*, P2*, and etc. are less accurate than P1, P2, and etc. were. This is another way The Process can leave us worse off than we started.

Again, as a minimal anti-skeptical commitment, it is natural to say that in some sense failure due to irrationality is the unusual case. Even if we normally fail to be maximally rational, we usually get close enough that The Process still leaves us at least better off than we were when we started. And this fact seems like it ought to have something to do with why the mere possibility of failure due to irrationality is compatible with a default presumption in favor of The Process. These remarks are cursory and even in this vague form controversial; it would be the task of a different paper to give them in a more perspicuous form. For present purposes just allow that there is some way, which we leave obscure, to describe the interaction of merely possible failure due to irrationality with the general case such that the general goodness of The Process can be affirmed.

So there are some default presumptions to the effect that for any arbitrary piece of evidence E, if we were to attempt to assimilate it by way of The Process, then we would thereby come to have more accurate beliefs; the default is against assuming that E is misleading or that our response will be irrational in an accuracy-destroying way. What I want to ask now is: what happens when that default assumption is disrupted? That is to say: what happens when we have strong evidence, for some piece of evidence E, that it is misleading, or that our evaluation of it would be irrational?

Let's look at some cases. So, for instance, imagine that you're me in one of the above two scenarios. But this time, rather than being caught unaware, someone tells you beforehand: the radio station's getting pranked today. Someone's going to tell them that there's a nuclear war on! Or, alternately: that stuff you're always doing with clouds doesn't make a whole lot of sense. You know clouds are just vapor, right? If they are sufficiently persuasive, you may then come to agree.

If you are so convinced, then when you hear the radio report, or you see the ominous clouds, you will ignore them. If one becomes convinced beforehand that some piece of evidence will be misleading, or that the response it provokes will be irrational, then that can in turn rationally change the way that one responds to that evidence. So long as one has the power to ignore, there is no threat that one will be somehow forced to make what one foresees as a bad epistemic move. We may say: so long as one has the power to ignore, one never really acquires evidence that one *will*, by way of some piece of evidence E, update one's beliefs to become less accurate. At most one can have evidence that one *would have* updated one's beliefs with E in such a way as to become less accurate—*would have* had one not known this very fact. But since one does know this fact, one instead just ignores E.

Unfortunately, not all cases are so easy. Sometimes we may rationally become convinced beforehand not only that some evidence E will be misleading, or that some response to E will be irrational, but that we cannot avoid that result by ignoring E. This because we cannot ignore it—if we acquire it, we simply *will* respond in a way so as to degrade our accuracy. One example is information about e.g. race and gender. Suppose the evidence in question contains facts about someone's race or gender. I take it that social scientists have provided definitive evidence that, as a matter of psychological fact, most people cannot simply ignore facts about race and gender. Rather, race and gender influence all sorts of evidential assessments whether we'd like to ignore it or not.

What to be done then? If we cannot ignore the evidence, and we know ahead of time that when encountered it will have deleterious effect, we need to do our best not to get it; to forget it when we get it; to prevent ourselves from being in a position to act on it after we get it, and so on. That sounds odd in the abstract, but the mechanisms by which it can be carried out are not so unfamiliar. Suppose that I am hiring a new member for my chamber orchestra. I may choose to listen to the auditions from behind a screen that obscures the tryouts—my purpose in doing so is to prevent myself from learning the races and genders of the auditioners. I am aware that implicit biases can cause people to misevaluate musical talent. So in order to prevent my assessment from being so corrupted, I make sure to not to learn certain facts about the people I'm hiring.[57]

In the racial case we have strong evidence that race influences our unconscious responses in a way that bends away from accuracy and toward stereotype. I do not think, however, that failures due to irrationality connected to un-ignorable evidence are confined to such largely unconscious evaluations. Rather, we can have strong evidence that our explicit, fully articulated, and fully conscious judgements in response to some future piece of evidence will involve a failure due to irrationality—even knowing full well and ahead of time that this is so.

To illustrate: suppose I have spent many years in an abusive relationship. I have finally found an opportunity to walk out, but have left some prized possessions behind at my former boyfriend's place. I want to get them back, but I'm afraid that if I return I'll see him. And I'm afraid that if I do he'll talk me into taking him back. I am extremely sure that leaving him is the right decision, but at the same time I know that he has, over the years, developed power over me, and that he understands how to break down my resistance.

Take the evidence here to be the content of the speech my ex-boyfriend's is planning to give me. Here I have strong evidence that, when it comes to this particular piece of evidence, that if I encounter it then The Process will fail due to irrationality. I have evidence that even though nothing he says will present an actual reason for me to return, I will nonetheless be convinced by his speech to return. And this even if I go into the situation with a steely resolve, if I tell myself beforehand not to be fooled by him again, and so on, and even though the reasoning that I anticipate being faulty is of the fully explicit sort that we ordinarily classify as under our direct control. So

[57] This adoption of this screened audition process in major orchestras has coincided with a substantial growth in their proportion of female musicians. For an analysis suggesting that a significant portion of this growth is indeed attributable to the adoption of the screened auditions, see (Goldin and Rouse 2000).

again, what I ought to do is: make sure I do not give my ex a chance to talk to me. I ought to send a friend to get my things, or perhaps I ought to give up on them entirely.[58]

The Process is generally good, and surely this is part of the rationale for our general desire to base our beliefs in the most evidence possible. But focusing too hard on the general goodness of The Process can lead us to mistakenly make the absolute claim that for *any* piece of evidence, we should *always* want to acquire and take account of it. Attention to fairly ordinary cases can help dispel this thought. More evidence may usually be better, sure, but all the same there are cases where our attitude toward a potential body of new evidence is decidedly negative. In these cases, we may well recognize the existence of some evidence while nonetheless wanting very much not to acquire it; if we acquire it, we may not want to consider it; if we consider it, we may not want to base our beliefs off it; and so on. I take it that the cases in this section fit that pattern.

[58] There is something of a parallel here between these claims and some made in the literature on overall practical obligation. I say: when you anticipate you *will* respond irrationally to evidence, you should avoid or ignore the evidence. Actualists about practical obligation say: when you anticipate you *will* act badly in a choice situation, you should avoid the choice. These claims seem close, and similar intuitive cases can be made to tell in favor of both. Possibilists, though, say: what you *should* do is not avoid the choice—rather, you should encounter the choice and act well (even if this is not what you anticipate will happen, it is all the same what you *should* do). Given that my position looks like an epistemic analogue of actualism, and actualism opposes possibilism, does that mean that possibilists ought to reject my position? Although there is a parallel, I'd suggest the two lines can be pulled apart. Here is one disanalogy: in the cases which actualists and possibilists argue over, the subject knows what the practically obligatory choice is and the problem is that they anticipate they may be too weak willed to select it. By contrast, in the cases I discuss, the subject has no problem with epistemic continence. They anticipate believing that which they take to be well-supported, they just think they will likely be wrong about what that is. This difference seems, at least to me, to be significant enough. So, there is at least some room to doubt that all salient facts about practical rationality port over to the theoretical case, and so at least room to think that my argument here does not require taking a controversial stand on the issue. Thanks to an anonymous reviewer for calling this parallel to my attention.

4.3 PEER DISAGREEMENT AND THE CONCILIATORY VIEW

The social-scientific data on racial bias presents strong evidence that one will respond irrationally to certain socially loaded categories, like race and gender. And one’s absolutely catastrophic history of decision-making with regards to one’s ex-boyfriend may present strong evidence that one will respond irrationally to his entreaties. Those are two fairly particular instances where we have rational expectations of future failure. Are there any more general classes that we can fruitfully analyze?

Yes. In particular, in this paper I will be focused on the philosophical treatment of disagreement. Broadly: when we learn that those around us who we take to be reasonable and well-informed have come to differing conclusions, that information may create a rational expectation of failure.

More specifically: in treating disagreement I will focus on one extremely special case—the case where one finds oneself disagreeing with someone who one takes to be, in a specific sense, an ‘epistemic peer.’ I make this choice because there already exists a literature on this case, and because its artificiality makes it tractable. I think the verdict I reach will generalize broadly; others may disagree. Regardless, before we can worry about the proper generalization we need to get the case itself right—and given the many divergent treatments it has received, getting straight on that is already a big enough task for this paper.

The special case is that one holds the following conditional credence: for some person A, and some proposition P, conditional on one’s disagreeing with A over P, one’s credence in P is ½.

As an example where such a credence would arise, consider *Two Bright Students:* Misty and Ash are good friends. They are also both clearly the best students in their class. Their averages have hovered around the same stellar number throughout the course so far, and they have traded the top class rank back and forth. In the past, when they have disagreed over the answer to a problem, they have each been right an equal portion of the time. Their final exam, rapidly approaching, consists in a single true or false question. This question will be neither so easy that they can just 'see' the answer, nor so difficult that they are just hopeless at figuring it out: it lies in the broad range of questions that they have been tackling with roughly equal success over the course of the semester.

I claim that: in the above case, it is rationally required for Misty to think of Ash that if, on the final exam, Ash disagrees with her over whether the answer to the final question is true, then it's 50 / 50 which of them is right—and, by extension, it's 50 / 50 whether the proposition in question is true. So the story above is one on which Misty is rationally required to hold of Ash the special sort of belief I'm interested in.[59]

Now suppose it is *A Frustrating Final:* Misty and Ash take the final exam. As they walk out, Ash turns to Misty and says "I got T." Misty replies: "uh oh, I got F."

Having learned of Ash's disagreement, what should Misty think?

The answer may look trivial: she should think it's 50 / 50! If Misty had the conditional credence described, doesn't it just follow that she should update according to it? It does not. Even supposing that conditionalization is the correct general way to update one's beliefs, it's still the

[59] It's worth noting that I identify conciliationism, the position I am concerned to defend, with a particular way of handling this special belief. But there are people who can be conciliationists in that sense while still retaining what are intuitively 'egoistic' and/or 'nonconformist' overall views of disagreement—as, for instance, if they thought it *a priori* almost impossible to get into the situation of regarding someone else as your peer in the specified sense. See for instance (Schafer forthcoming), and as noted in footnote no. 64 possibly (Wedgwood 2010). I think those views can be ruled out, but doing so requires a separate argument from the one given here. It is also worth noting that this is not the only way of defining what it is to be an 'epistemic peer.' That's fine; my thesis should be understood just as the claim that epistemic peerhood *understood in this way* requires a particular response. I take no stand on what other notions of peerhood require. See e.g. Lam's (2011) and (2013) for examinations of how varying the understanding of peerhood in play can correspondingly vary the performance of different disagreement policies, and see footnote no. 63 for further discussion of Lam's work.

case that one ought conditionalize on the *strongest* evidential proposition one learned. Misty's earlier belief was conditional on her and Ash disagreeing over the final question. Over the course of the story told, is 'the final exam happened and Ash and I disagreed over the final question'' the *strongest* relevant thing she learned?

Call this fact 'the mere fact of disagreement.' The question then becomes whether, during the course of taking the final and then comparing answers with Ash, Misty learned anything epistemically relevant over and above the mere fact of disagreement. If the answer is 'something important' then Misty may need to change her mind. She may then no longer think it's 50 / 50.

Of course, Misty may notice that Ash is visibly drunk when he's taking the test, or alternately that he has cheated and smuggled in the answers, or whatever, and thereby come to be very sure that he's either right or wrong. That would be fine so far as it goes, but it's not what we're interested in. Rather, the question is about whether there's anything learned in cases of disagreement *per se* over and above the mere fact of disagreement.

So there is a substantive question here. I will argue that the answer to that substantive question is the one that you might have thought was trivial: Misty should still think it's 50 / 50. If she learns anything over and above the mere fact of disagreement, it's not the sort of thing that can rationalize a change in confidence. Call this view the conciliatory view.[60]

[60] This view is popular in the literature, and leading defenders include Christensen (2007), (2009), (2011), and Elga (2007), (2010a).

4.4 INDEPENDENCE AND EXTRA WEIGHT

What might Misty learn in the course of the disagreement, over and above the mere fact of disagreement? Well, she sees the question itself. She reasons through it. She arrives at her answer. Beforehand, she was agnostic about who would be right if Ash disagreed with her on the final question, whatever that was. But now she sees that Ash is disagreeing with her not just on a question, but on *this* question, and that he is doing so by getting *that* answer.

How does this change things? Suppose Misty reasons as in the following way, which I'll call *D*: I got T on the final question. And T is the answer to the final question. Ash got F. Since T is the right answer, Ash's answer is wrong. Since Ash got the wrong answer, he must have reasoned incorrectly in this case. Since I can now see that he reasoned incorrectly, our situation is no longer symmetric. I no longer take him to be my peer when it comes to this specific question, and if I need take any account of his opinion at all it need not be as much as I antecedently would have thought.

In reasoning in this way, Misty leverages the very facts under dispute in the disagreement while considering how she ought to respond to it. If that's legitimate, then it will never be the case that thinking it was 50/50 in advance of the disagreement requires thinking it continues to be 50/50 afterward. Rather, one will always discover something new, something capable of disturbing one's original assessment: namely, the answer one got. And so conciliationism would be false.

The reasoning in D, however, can seem problematic precisely by virtue of its appeal to the very facts under dispute. To block such reasoning, conciliationists have proposed *Independence:* in evaluating the epistemic credentials of another person's belief about P, in order to determine how (if at all) to modify one's own belief about P, one should do so in a way that is independent of the reasoning behind one's own initial belief about P.[61]

If Independence is true, then reasoning as in D is illegitimate. And if reasoning as in D is blocked, then it is plausible that there is nothing relevant that Misty learns over the course of the disagreement to disrupt her initial assessment. In other words, conciliationism looks to follow. Both ways, conciliationism seems to stand or fall with Independence.

Ought we adopt Independence, and by doing so block reasoning like that in D? One reason to think that we should is that the reasoning in D looks dogmatic, and as such epistemically unimpressive. This negative first impression can be made more precise. Suppose we interpret reasoning as in D as licensing Misty to, in the course of disagreement, give preferential treatment to her own convictions over Ash's merely on the basis that they are her own—this is at least one way to fill out the dogmatic undercurrent to D. Since this interpretation of D involves giving one's own view extra weight, call this view *the extra weight view*.

Elga has argued against the extra weight view by pointing out that it appears to allow Misty to 'bootstrap' her way into undeserved confidence in her own abilities; since such bootstrapping is epistemically abhorrent so too must be views which permit it.

How does the bootstrapping objection proceed? Elga offers the following reductio: suppose it really were permissible to give one's own view extra weight. If that were so, then after disagreeing with a peer, one could be rationally confident that one was right; but if *that* were so, then one could also become rationally confident in the propositions one's having been right entails, namely that one has a better track-record than one's friend, and hence that one is more reliable

[61] This principle is proposed in e.g. (Christensen 2007: 16-17) and (Christensen 2011: 1-2). Independence, as formulated here, does *not* block off making reference to the existence or even the properties of one's reasoning—again, provided that those properties are picked out in a way that does not presuppose the correctness of that reasoning.

after all—and all this would be possible in advance of actually receiving any independent confirmation of one's rightness. If this were really rational, that must be because disagreement with a peer is evidence that one is the better judge. But that's absurd. The mere fact of disagreement, absent any independent confirmation that one is right, is no evidence at all that one is the better judge. Hence, the extra weight view is false.[62]

This line of objection is raised against views which would allot oneself extra weight, but it applies just as much to views which would allot one's interlocutor extra weight: there the absurd conclusion is just the reverse of the one derived before, namely, that one can take the mere fact of disagreement to be evidence that one's interlocutor is more reliable. The point here is just that the existence of disagreement on its own is no evidence at all either way. Therefore, Elga's bootstrapping objection aims to rule out any deviation from the conciliatory view whatsoever.

As it stands, though, this argument does not do much to motivate Independence. The bootstrapping argument assumes that the mere fact of disagreement is the only epistemically relevant thing learned and then, on that basis, constructs a dilemma where the choices are conciliation or extra weight. And it's true that if that were the dilemma then conciliation would be thereby established as only sensible option. But the assumption which generates the dilemma, namely that the mere fact of disagreement is the only epistemically relevant thing learned, is something that itself needs to be demonstrated.

Indeed, non-conciliatory theorists have responded to the bootstrapping argument by allowing that it is fatal to the extra weight view—where that view is the one that takes the mere fact of disagreement to be evidence—but then carefully distinguishing their own views. They have done so by giving an account of the evidence they take to arise over the course of disagreement,

[62] This argument is a paraphrase of (Elga 2007: 12-15).

over and above the mere fact of disagreement, and describing it in such a way that it seems to be the sort of thing that could rationalize a change in view. These replies can successfully defuse the objection as it is phrased above: there it takes the form of accusing non-conciliatory views of taking the mere fact of disagreement to be evidence when it is not, in fact, really evidence. When non-conciliatory theorists give an account of the more robust evidence that they take to arise in disagreement such that it looks like it is the right sort of thing to do the work required, then they adequately respond to the objection.

I will argue, however, that there is a closely related problem which has not been the subject of so much attention. Elga's objection can be reformulated not in terms of the evidence that there is, but in terms of how one knows one will respond to it. Even though non-conciliatory views can give a richer description of the evidence at hand than the extra weight view does, they nonetheless bear a problematic relationship to it: namely, we can ascertain ahead of time that if one tries to follow a non-conciliatory view, one will in fact reach the same results that one would have reached by actually following the extra weight view. But following the extra weight view looks like a bad deal from the perspective of accuracy—and since accuracy is a matter of what you believe, not why you believe it, if the extra weight view looks bad from the perspective of accuracy then so too do all the views the attempted following of which would yield the same results.[63]

[63] My approach here bears some notable similarities to the line advanced by Barry Lam in his (2011) and (2013). He seeks to inform the normative debate over peer disagreement by answering closely related non-normative questions about how different views will perform under application; I try to inform the normative debate over peer disagreement by answering non-normative questions about how agents expect their views to perform under application. I take it that the most significant difference in our approaches is that we focus on different understandings of what it is to be a peer—I understand peerhood in terms of the holding of a conditional credence of the type outlined in section 4.3, whereas he understands peerhood in terms of agents equally well satisfying a measure epistemic success (he considers multiple such measures). These different operative understandings of peerhood lead to significant downstream differences in our conclusions. It is interesting to sort out which of these different conclusions represent actual disagreements and which simply answer to different questions; my sense is that there is much of the latter, though I can't fully trace that out here.

Before I give this argument in greater depth, I need to introduce the non-conciliatory views I am to criticize. I will canvass three major alternatives. First I'll describe Enoch's steadfast view (Enoch 2010).[64] Then I'll describe Kelly's total evidence view (Kelly 2010), (Kelly 2013).[65] Finally, I'll describe the right reasons view; since this last is in all relevant structural features just a variant of the total evidence view, my description will be quite brief. The point will be to get onto the table their conceptions of the relevant evidence that arises over the course of disagreement. Once that's done, I can go on to argue that, as so specified, that evidence bears a troubling relationship to accuracy.

4.5 THE STEADFAST VIEW

First, what is Enoch's position? He argues as follows. The only way in which it is possible to form judgments about who is a peer and who is not is to look to someone's track record: how reliable has she been on this subject in the past? Furthermore, the only way to do that is to compare her history of judgments to one's own and see how well they match—we have nowhere else to start, when judging the reliability of our peers, than from our own views. But if we methodologically

[64] By contrast to Enoch, it is unclear to me exactly how to categorize Wedgwood's view with respect to conciliation. In his (2010: 237) he seems to endorse Independence, while claiming that it will be hardly ever applicable, because one hardly ever ought to have the conditional credence that characterizes peerhood beliefs as I've defined them. Yet later (2010: 243) he seems to endorse reasoning relevantly like D. I suspect this might be because he takes it to be the case that one typically lacks any relevant credence about one's peer's reliability prior to some disagreements, and hence reasoning like D need not conflict with Independence. This would, in my locution, make him a conciliationist, albeit one who does not think conciliationism is a particularly interesting or widely applicable thesis.

[65] One significant view that I do not explicitly discuss here is Lackey's "justificationist" view (Lackey 2008a), (Lackey 2008b). I take the justificationist view to be, in the respects relevant to the argument of this paper, the same as the total evidence view: the justificationist's reliance on one's *actual* initial degree of justification, as well as on the *actual* propriety of one's mental goings on as encoded in 'personal information', will together have the same effect as the total evidence view's reliance on the *actual* disposition of the non-psychological evidence. So, I take my criticisms of the total evidence view to apply *mutatis mutandis* to the justificationist view.

take ourselves to be right in our evaluation of putative peer's past reliability, then it be arbitrary not to do the same with our present conflict. Thus, Enoch advises, we ought respond to peer disagreement by, among other things, demoting our putative peer to some degree; her new track record, this judgment included, is worse than their old one. So Independence is false, and the reasoning in D is legitimate.

But doesn't this involve treating the mere fact of disagreement as evidence that one is a better judge that one's peer? It does not. Enoch is careful in differentiating his view from the extra weight view: he insists that the grounds on which we demote our peer is *not* the mere fact of disagreement, but rather, it is the deterioration of their track record. And the deterioration of their track record *is* good evidence that they are not good judges (Enoch 2010: 981-986).

To illustrate: suppose I take you to be my peer on matters p-related. I judge p. I then discover that you have judged not-p. On Enoch's view it is appropriate for me to demote you, but not on the grounds *that I judged p* and *that you judged not-p*, but rather, on the grounds *that p* and *that you judged not-p*. Of course, what makes it the case that it's appropriate for me to demote you on the grounds *that p* is that I have judged p, but that does not make judging p itself my grounds. Our beliefs are, and necessarily must be, transparent insofar as we act from their contents *directly*, rather than from hedging propositions like "I judge p."[66]

Since I am acting directly from the content of my judgment, rather than from the fact that I so judged, my judgment is based in your deteriorating track record itself rather than merely the fact that I so judged. There is some relevant evidence here, the evidence of track records, and that distinguishes the steadfast view from the extra weight view.

[66] C.f. (Wedgwood 2010: 242-243), which both makes this same point and takes it to be central to the epistemology of disagreement, which is why it is tempting to classify Wedgwood as steadfast despite his aforementioned seeming-endorsement of Independence in footnote no. 64.

4.6 THE TOTAL EVIDECE VIEW

In cases of peer disagreement, Independence screens off the particular belief in question—thus prohibiting "p, therefore you're wrong about ~p"—but that's not all it screens off. It also screens off the reasoning used to arrive at p. There are obvious reasons for this. Imagine p was screened off but the reasoning supporting it was not. Then one could simply re-conclude p from that reasoning, and then proceed as in D. But Independence is designed to block D. So Independence must screen off that supporting reasoning as well.[67]

This, however, opens up Independence to the charge that it throws out evidence. After all, if I was correct in my reasoning from my evidence to p, then that evidence really does support p. I should not ignore it when coming to my final opinion. Conciliation falsely treats the case as if the only evidence available in the wake of a disagreement with a peer were our differing beliefs. But there is more evidence than that—there is the evidence on which we based our differing judgments in the first place.

So, for instance, Kelly imagines a case of peer disagreement where we both initially form our views on the basis of some evidence E. After we consult with each other, and discover our differing conclusions, our new total evidence includes both the original evidence E and the fact that we reached contrary responses; call this new pool of evidence E*. He then puts the point as follows: "Notice that, on the Equal Weight View, the bearing of E on H turns out to be completely irrelevant to the bearing of E* on H. In effect, what it is reasonable for you and I to believe about H... supervenes on how you and I respond to E... E gets completely swamped by purely

[67] (Christensen 2011: 18) makes this point while introducing a problem about how to define the scope of Independence.

psychological facts about what you and I believe... But why should the normative significance of E completely vanish in this way?" (Kelly 2010: 124). The original evidence, E, is still available to us, and so it should not vanish.

But what of that original motivation for Independence—ruling out D? Would Kelly assent to "p; you think ~p; therefore you are a worse judge of p than I?" The answer, it turns out, is 'it depends.' If, after adding the fact that you think ~p to my total evidence, it still supports p on the balance, then I am allowed to so proceed. If it does not, I am not (Kelly 2013: 43-45). The philosophical assumption that there is any general epistemic rule that can be decided beforehand is false. Of course, deciding what the total evidence supports in any particular case is hard, and certainly more difficult than simply applying a formal rule—but "such are the burdens of judgment" (Kelly 2013: 52).

Kelly's view, then, is also distinct from the extra weight view. Whereas the extra weight view tells you to assign extra weight to your own answer as such, Kelly's view tells you to attend to the total weight of the evidence; if the evidence really does favor the answer you initially came to, then you should respond to it by retaining a higher credence in your own answer. But if it favors something else, you ought to believe that other thing. What is important here is just the disposition of the evidence, and nothing in that description makes reference to your view as such.

4.7 THE RIGHT REASON VIEW

The total evidence view says: conciliatory views overlook the non-psychological evidence. When Ash and Misty disagree about the answer to the final exam, the facts that each holds the view that they do may constitute some symmetrically balanced evidence in favor of each of their answers.

But there is the further non-psychological evidence to consider and it may well break the symmetry between them.

The right reasons view expresses the same core idea in more extreme form. Right reasons theorists hold that not only is there some relevant non-psychological evidence, but that the psychological evidence never matters as such. Ash and Misty should both just believe whatever the non-psychological evidence supports, and that remains the same no matter how many people's dissenting voices they are exposed to.[68]

4.8 NON-CONCILIATORY VIEWS FUNCTIONALLY ENTAIL EXTRA WEIGHT

So: all of these non-conciliatory views posit something epistemically relevant which comes into Misty's possession over the course of the disagreement. That may be a reason based in a changing track record, or it may be some combination of the evidence on the final and then the discovery of Ash's contrary verdict, or it may be entirely the evidence on the final. Call this epistemically relevant stuff, whatever it is, E. The non-conciliatory views say: E is the sort of thing that rationalizes deviating from the 50/50 verdict with which Misty began. So non-conciliatory views have a story about why both Independence and the conciliatory views that rely on it are false. Independence would exclude, and conciliatory views ignore, this E.

[68] Right reasons theorists may allow that there are other propositions for which the psychological evidence provided by disagreement matters—for instance, that all the students answered T may be evidence that it's rational to conclude T. The distinct commitment, though, is that it is nonetheless not (in this case) evidence that T. For more on the possibility 'level splitting' see e.g. (Weatherson 2013), (Weatherson ms.), and some passages in Kelly's earlier (2005) view. For a defense of right reason without level splitting see (Titelbaum 2015).

But forget for a moment what E rationalizes. That is to say: forget what Misty *should* do with it. It's worth asking instead what Misty *will* do with it, in the course of trying to do what she should. And on this, proponents of the non-conciliatory views we have considered agree. Misty will respond to E by becoming more confident that she, rather than Ash, has had the better of the disagreement.

We can introduce the following bit of terminology: let's say that epistemic view A *functionally entails* epistemic view B iff someone who attempts to act in conformance with A will, in fact, wind up arriving at all the same attitudes as someone who was actually acting in conformance with B. What non-conciliatory theorists concede—and it is hard to see how they could not—is that their views all functionally entail the extra weight view.

This is easiest to see with respect to the steadfast view. On the steadfast view, both Misty and Ash receive an E such that each of them may become rationally confident in their own correctness on its basis. But this is just the same result that would be arrived at by parties following the extra weight view. As such, the steadfast view functionally entails the equal weight view.

Enoch acknowledges this but is untroubled. He says: one can foresee, on his view, that one will arrive at all the same verdicts as the person who favored their own view with extra weight, but nonetheless the person following his view does not act under the intention of giving herself extra weight as such, and this difference between foresight and intention is epistemically relevant (Enoch 2010: 989-990). So he acknowledges the functional entailment while disputing its significance. But that's fine; the point for now is just that there is such a functional entailment.

The total evidence view is not symmetrical in the way that the steadfast view is. On the total evidence view, Misty and Ash receive a single E, and that E is such that it is tilted toward the person who was in fact better responding to the evidence when they formed their initial judgment.

So both of them ought to arrive at the same judgment, namely one tilted toward the person who was in fact better responding to the evidence when they formed their initial judgment. Does this lack of symmetry functionally distinguish it from the extra weight view?

It does not. Since the answer to the question of which party E tilts toward is itself dependent on who was right in their initial judgment, in attempting to ascertain and thereby properly respond to the tilt of E both parties will have no alternative to re-deploying that same initial judgment. As Kelly puts it: there is no 'warning bell' that goes off when you are misevaluating the evidence, and which lets you know that you're the one in the dispute who ought to lay down your arms; furthermore, the relevant facts may not seem to be facts at all from your perspective (Kelly 2010: 165, 167-172). So in trying to take account of the tilt of E, it's inevitable that one will de facto take it to tilt toward oneself. The total evidence view, just as much as the steadfast view, also functionally entails the extra weight view.[69]

The same is true of the right reasons view. On that view E does not merely tilt toward but is rather fully in line with whichever party was right in their initial judgment. So, for the same reasons, it will also functionally entail the extra weight view. The difference is just that it will functionally entail a particularly weighting: namely, the extremal weighting that places all weight on one's own view and none on any interlocutors.

So, non-conciliatory views all functionally entail the extra weight view. When we forget about what people *should* do and focus just on what they *will* do, we see: people with non-

[69] Indeed, this is why (Setiya, 2012) suggests, in friendly development of the total evidence view, that although the stronger status of justification should be reserved only for the correct party, the status of blamelessness should still be allowed to the incorrect. The incorrect party is trying to follow the correct epistemic norm—it's just that their situation is hopeless, and so blameless. As with Enoch, this again acknowledges the functional entailment while holding out that there is a significant non-functional epistemic difference; here the difference between mere blamelessness and actual justification.

conciliatory views will respond to the evidence E that arises in disagreement by acquiring just the same attitudes that one would acquire by putting some extra weight on one’s own view, taken as such.

Given that some there is some, how *much* extra weight exactly is functionally entailed by these views? Misty starts out, before the final, taking things to be 50 / 50. How far away from that will she get, on the basis of E? Let’s name that distance, whatever it is, N. Misty gets up to 50 + N / 50 - N.

In introducing ‘N,’ I should flag what I do and do not assume about it. My assumptions are minimal. So: I make no assumption that there is some N constant across these views—as already noted, what’s distinctive about the right reasons view is it, unlike the other views discussed, seems to functionally entail a very high N. Nor do I assume that N will be constant across different developments of the individual views; you could for instance have different steadfast views that varied in how stalwart they commanded one to be. Nor do I assume that for each individual view it will assign some single N constant across different types and instances of disagreement: maybe N will depend in each instance on all sorts of variegated local facts. Nor do I assume that the actual value N will take for any given disagreement can be precisely deduced in advance.

There are, however, two things that I *do* assume. The first thing I assume is that N is sometimes non-trivial. Non-conciliatory views would not present a very interesting or exciting alternative if they posited some E such that it allowed Misty to be as much as an extra thousandth of a percent more confident in her correctness. Rather, the difference ought to be sometimes substantial. The second thing I’ll assume is that, even if the value of N is not precisely knowable in advance of any particular disagreement, nonetheless a reflective and rational subject could at least estimate it.[70]

4.9 EXTRA WEIGHT DECREASES EXPECTED ACCURACY

So, the non-conciliatory views give characterizations of Misty's evidential situation on which they need not say that she changes her mind in response to the mere discovery that Ash disagrees. Rather, she changes her mind in response to E, evidence that she acquires along the way and which is described so as to look like the right sort of thing to rationalize her change of mind. Non-conciliatory views thereby distinguish their rational structure from the rational structure of the extra weight view. Nonetheless, they admit a functional entailment from their positions to the extra weight view. In this section I look harder at the consequences of that admission.

The question to ask now is: given that E is as described, how should Misty see it as figuring into The Process? And the objection is: given Misty's peerhood belief, and given the non-conciliatory views' functional entailment of the extra weight view, Misty should think that undergoing The Process with respect to E will reduce, rather than enhance, her accuracy.

So, let's imagine that Misty is preparing to head off to her Frustrating Final. Misty is a non-conciliationist, and she knows it. She also knows that she and Ash are about to take a test with a

[70] On some views rational subjects will have perfectly sharp credences in even the most outlandish of propositions: see, for instance (Elga 2010b). That would be congenial to my argument, but I need not assume it. All I need to assume is that: however well behaved things need to be to yield an estimate that is itself well behaved enough to figure in practical reasoning, N is at least that well behaved. This is satisfiable both if N is poorly behaved but good behavior is not required, or if good behavior is required but N turns out to be well behaved.

single true or false question, and that they will compare their answers afterward. What are the relevant possible outcomes, and what does Misty now believe are the chances that each will occur?

There are two relevant branch points, and so three relevant outcomes. The first branch is that Misty and Ash may either get the same answer or they may get different answers. The second branch is that, provided they get different answers, Misty can be right and Ash wrong or Ash right and Misty wrong.

What are the chances on each horn of the first branch, as Misty now sees things? The chance that Misty and Ash will agree depends on each of their independent reliabilities. Let 1-M be the chance that they agree, and hence M the chance that they disagree.

What are the chances on each horn of the second branch, as Misty now sees things? The second branch is over, should they disagree, which of the two is right. That one is straightforward: Misty takes that to be 50 / 50. That's just the content of her peerhood belief as we have defined it.

That's what she thinks about whether she will be right. What does she think about what she will wind up believing? Misty knows that, should they disagree, she will wind up believing that she is 50 + N% likely to be right; such is the foreseeable functional entailment we outlined in section 4.8.

All together then: Misty now believes that if they disagree it's 50 / 50 whether she or Ash will be right. But she also believes that if they do disagree, then in acquiring and responding to E she *will* become convinced that it's 50 + N% / 50 - N%, favoring her. Should she foresee that change producing an increase or a decrease in her expected accuracy? As I indicated earlier, my goal in this section is to argue that the answer is: she should foresee a decrease. I will defend that claim in two ways. First, with a lightly technical gloss, I'll note that it follows from the use of suitable accuracy measures. Then I will go on to offer some informal analogies which reinforce that verdict.

So, first: how should we try to quantify Misty's expectation? If she does in fact wind up being right, then the increase of N will wind up making her credence more accurate; if, on the other hand, she winds up being wrong, then the increase of N will wind up making her credence less accurate. Since she now thinks it's 50/50 which she'll get, whether this change has a positive expected value for Misty depends on whether the gain in accuracy that she gets in the good case is larger or smaller than the loss in accuracy that she undergoes in the bad case.

This will depend, in turn, on what 'scoring rule' we use to measure accuracy. In general: we want to say that credences which are farther from the truth are worse than credences which are closer to it. A scoring rule is a mathematical tool that tells us, given our credence and the actual truth, how to measure that distance between them, and thereby lets us draw the relevant comparisons.

Scoring rules can be divided into those which are 'strictly proper' and those which are not. A scoring rule is strictly proper iff it renders all consistent credences self-endorsing with respect to expected accuracy. That is to say that, once we use some credence in setting the odds for an expected value calculation, the calculation then selects that very same credence as the one which uniquely maximizes expected accuracy. But notice that is just what we are asking about here: given that Misty takes it to be 50 / 50 who will be right, how accurate does she expect her later 50 + N / 50 - N credence to be? Adopting a strictly proper scoring rule settles the answer as: 'less, because given that we are using 50 / 50 to set the odds for the expected value calculation, it will follow that

50 / 50 uniquely maximizes expected accuracy.’ *Ipso facto*, 50 + N / 50 - N has a lower expected accuracy.[71]

So given that our scoring rule is strictly proper, we can get the conclusion we set out to: Misty will see acquiring and responding to E as lowering her expected accuracy.

One might worry that by appealing to strictly proper scoring rules I rest my conclusions on an arbitrary choice in formalization. I have two responses: the first response is that the choice to use a proper scoring rule is not arbitrary, and in fact expresses no new commitment over and above those already introduced. In sections 4.1 and 4.2 I offered, as an anodyne starting point, the observation that we expect undergoing the Process to improve our accuracy, absent special reason to think otherwise, and that getting that increase in accuracy is part of the reason why we are concerned with acquiring new evidence. For this to be generally true, we already need to be committed to using strictly proper scoring rules.[72] The second response is that strictly proper

[71] So, for instance, take the Brier score, which is strictly proper. It measures accuracy indirectly by way of measuring inaccuracy (which we then minimize in order to become more accurate). To calculate the inaccuracy of a credence, it squares the distance between it and the actual value (0 if false, 1 if true). So, for a credence of .5, its inaccuracy will always be (1 - .5)^2 = (0 - .5)^2 = .25. A credence of .5 + N will have either inaccuracy (1 - (.5 + N))^2 or (0 - (.5 + N))^2, depending on whether the proposition is true or false, respectively. If it’s taken to be 50/50 which, then the expected inaccuracy is .25 + N^2, and so this gain of N confidence would lead to an expected increase of N^2 inaccuracy.

[72] To illustrate, consider a widely discussed improper scoring rule: the linear scoring rule. Under the linear scoring rule, the only credences which self-endorse from the perspective of accuracy are either maximally opinionated or perfectly indifferent. As it turns out, if someone follows the linear scoring rule, then any time they become less opinionated (without becoming perfectly indifferent), then by their lights both before and after the change, they expect their new accuracy to be lower than their old one. But surely sometimes it is rational to lose marginal confidence, as, for instance, when the new jobs report leads me to go from being .7 confident that the President will be re-elected to merely .6 confident. If the linear scoring rule is correct, there are going to be all sorts of rational belief changes such that the person undergoing them nonetheless anticipates their accuracy falling.

The linear scoring rule presents a particularly dramatic example, but what goes for it also goes for improper scoring rules generally. Improper scoring rules force us to make a choice: either hold that considerations of accuracy are just normatively irrelevant, and so it is no problem that we sometimes expect ourselves to be doing poorly by our own accuracy-lights; or, alternately, allow that reflection on the nature of the satisfaction relation between probabilistic beliefs and the world actually substantially constrains the otherwise consistent beliefs we can reasonably hold and reason to—as, for instance, by ruling out all beliefs which are neither maximally opinionated nor maximally indifferent.

The first option is ruled out by the opening remarks of the paper—I take it to be platitudinous that we take ourselves to ordinarily be improving our accuracy, and that doing poorly along our lights accuracy-wise is a problem. The second option is independently implausible. Reflection on the satisfaction relation may reveal that inconsistent scoring rules are useful to quantify my objection to non-conciliatory view, but that their only feature I‘ve actually appealed to here—that random changes a fixed distance toward or away from the truth look bad epistemically—is anyway independently plausible.[73]

So, on reasonable ways of measuring accuracy Misty should foresee a decrease in her expected accuracy. That is the first part of my argument. The second part reinforces that verdict by turning away from the abstract question of scores and directly considering *what Misty's epistemic situation is like*, on these non-conciliatory views. In answering, we can buttress our earlier, more abstract claims by giving some comparisons that make it clear both what is going on in Misty's case and why she should anticipate that undergoing The Process with respect to E is a bad idea.

So: what is going on? On each family of view non-conciliatory view, Misty anticipates that undergoing the process with respect to E will fail. But, as it turns out, it is for different reasons in each case.

For the steadfast view, Misty anticipates the process failing due to misleading evidence. Whenever Misty anticipates a situation of peer disagreement, she thereby anticipates a situation where two pieces of evidence will be generated: one for her, one for the relevant Ash. These pieces of evidence will symmetrically support opposite conclusions about whose record has improved and whose has degraded. These Es by nature are generated in opposing pairs, and for each non-misleading E there is a paired misleading E.

beliefs are bad, but it is difficult to believe that it rules out whole swaths of consistent ones. C.f. (Greaves and Wallace: 2006: 17-18).

[73] So, for instance, Gibbard in his (2007) is skeptical of accuracy-first epistemology on the basis that the linear scoring rule seems to him to be a reasonable way of caring about truth, and it renders a hash of accuracy-first epistemology. Instead, he proposes that guidance value, rather than accuracy, is the appropriate candidate for the basic epistemic value. Although he thereby disagrees with the general methodological sympathies of the paper, his own view is nonetheless copacetic with the substance of the arguments, since his conception of guidance value also secures this feature.

Suppose that, instead of being ruled by a malevolent demon, we were ruled by an absurdist one. One day you run into the absurdist demon at a gallery opening and over free wine she lets it slip that she finds the idea of people coming to understand the world by way of reading newspapers to offend her sensibilities. And so what she does is this: every time some piece of non-misleading evidence is reported in a newspaper, she ensures that some other misleading newspaper report is also generated. The nature of this second report is that it always contains misleading evidence of the exact same strength as in the original report, and that this misleading evidence always indicates instead the opposite conclusion. Some newspaper reports evidence for P, which is true; she ensures that some other newspaper reports equivalently strong evidence for ~P, which, of course, is false.

If that happened, the first thing to think would be that opening the newspaper, reading the reports therein, and then believing them would lead to less accurate beliefs. It would be, for all you could tell ahead of time, random whether you were going to get one of the misleading or one of the non-misleading pieces of evidence, and random changes in one's beliefs are expected to degrade their accuracy.[74]

But the situation you would be in with respect to newspapers is just the same that we are all in with respect to disagreement, if the steadfast view is correct. When it comes to disagreement: for every non-misleading piece of evidence, there's an equal and opposite misleading piece, and antecedent reason to anticipate that you're as likely to get one as the other. And, unlike the newspaper case, there is no need for a demon to tip us off. It's an *a priori* feature of E, so described, that it must be this way.

For the total evidence and right reasons views, on the other hand, Misty anticipates the process failing due to irrationality rather than misleading evidence. For these views, there is only

[74] As above, this will be true under any strictly proper scoring rule.

a single E and it always points in the same direction as the correct assessment of the original evidence; there is no reason to think that it is systematically misleading. Yet there is nonetheless a reason to think that responding to it will systematically lower accuracy.

To see this, consider a new absurdist demon. This demon is such that her sensibilities are offended by your confidence in your rational faculties, and so she also has tampered with the newspapers. But this time, rather than mix in misleading evidence, she has instead ensured that all the reports are non-misleading. But she is still a villain, just more weirdly: when stocking the newspapers, for every piece of evidence that you will rationally appreciate she has also inserted elsewhere a piece of evidence that you will irrationally misjudge. She has looked into your soul and seen all your prejudices and incompetency, and he has thereby constructed a world full of good newspapers and good evidence—but yet nonetheless a world in which for each time some newspaper reports evidence for P such that you consequently appreciate P, another reports evidence for Q such that consequently you mistakenly conclude ~Q.

This absurdist demon is different than the first, but her effect on your relationship to newspapers is the same. Again, if this were revealed, the first thing to think would be that opening the newspaper, reading the reports therein, and then drawing conclusions on that basis would lead to less accurate beliefs. It would be, for all you could tell ahead of time, random whether you were going to react rationally or irrationally to the evidence therein, and random changes in one's beliefs are as before expected to degrade their accuracy.[75]

And again, the situation you would be in with respect to newspapers in this scenario is just the same situation that we are all in with respect to disagreement, if the total evidence or right reasons views are correct. When it comes to disagreement: for every piece of evidence you'll

[75] Again, this follows from using a strictly proper scoring rule.

rationally appreciate, there’s a piece that you’ll irrationally follow in an equal and opposite direction, and antecedent reason to think you’re as likely to get one as the other. And, again no need for a demon to tip us off. It’s an *a priori* feature of E, as so described, that it must be this way.

We opened this section by demonstrating that since all non-conciliatory views functionally entail the extra weight view, all non-conciliatory views thereby construe disagreements as situations wherein attempting the proper response degrades expected accuracy. And now we’ve explained, for each family of view, why that’s so. For steadfast views, it’s because they construe the evidence acquired during disagreement, E, as such that it by nature is generated in equal and opposite pairs. The result, then, is that in responding to it one anticipates failures of the process due to misleading evidence. For the total evidence and right reasons views, it’s because they construe the evidence acquired during disagreement, E, as such that by nature for every rational response it provokes it will also provoke an equal and opposite irrational one. The result, then, is that in responding to it one anticipates failures of the process due to irrationality. For both, and for either reason, the relation to accuracy is negative.

4.10 OBJECTING TO THE INFERENCE TO CONCILIATION

Allow that all this is so: non-conciliatory views functionally entail the extra weight view, the reason for this is that the evidence non-conciliatory views appeal to predictably prejudices oneself toward one’s own correctness, and the result is that one expects responding to that evidence to lower one’s accuracy.

But what follows? My preferred answer applies the conclusions we drew in section 4.2: since we can anticipate that trying to respond to E-type evidence leads to a decrease in our expected accuracy, we ought not respond to it. Instead, we ought to ignore it. And so, I think, we should be conciliationists. After all, once the E-type evidence is ignored, all that's left behind is the mere fact of disagreement, and all parties to the debate began by agreeing that *if* the mere fact of disagreement were the only epistemically relevant thing learned, then conciliation would be the only sensible response.

In this section, I want to get onto the table a way of resisting this conclusion. The response I'm going to consider accepts the earlier conclusions about the foreseeable consequences of trying to respond to E-type evidence, but disputes that we thereby ought to be conciliationists.

So far, I have treated *steadfast*, *total evidence*, and *right reason views* in parallel, because the functional equivalence to extra weight holds of all. But the considerations I'm about to advance are no longer indifferent among them. Rather, I am going to put aside steadfast views, as the response I am going to consider is available only to total evidence and right reason views. This response exploits the fact that these views conceive of E in such a way that it is biased toward the truth, and, as such, the functional equivalence with extra weight only holds because subjects expect that they will sometimes behave irrationally. Both total evidence and right reason views make use of such a bias toward the truth, and because of this I will refer to them here on out as BT views.[76]

The response has two parts. The first goes like this: it's true that you can anticipate that should you respond to some piece of evidence E, the credences you will form in so doing will be losers from the perspective of accuracy. But that is only because you anticipate that you may make a rational error, and, in general, the badness of the errors you make are going to outweigh the

[76] I borrow the phrase 'bias toward the truth' from Setiya's development of his (2012) non-conciliatory view.

goodness of the successes. But in those cases where you *don't* make a rational error, there is no argument that responding to E is bad; rather, the platitudes about the general goodness of responding to evidence should lead us to expect the opposite. Responding to E without making a rational error improves accuracy. And, as a proponent of a BT view, I have always thought that what you *should* do is respond to E *without* making any rational errors. No doubt has been cast on the goodness of *this* policy.

The second part continues: furthermore, not only is *successfully* following my norm (the only type of following I recommend) a good deal with respect to accuracy, but someone who is doing so may recognize themselves as so doing, and thereby need not worry about the general tendency of E-type evidence to lead to error. We may grant that Misty should expect that, at the level of description 'some item of E-type evidence,' responding to that item of evidence will lead her astray. But in A Frustrating Final she doesn't receive 'some item of E-type evidence.' She receives *this* item of E-type evidence. And she may correctly perceive that *this* item of evidence supports her. If she does so, she may maintain her belief that *in general* items of type E will lead her astray, but still take *this particular* item of evidence to support her correctness. So one can acknowledge the general claims about E-type evidence thus far argued while still following a BT view and fully endorsing the results they get by doing so. This because in each particular case, successfully following the rule will involve genuinely recognizing that they are getting it right.[77]

Now, of course, sometimes the people trying to follow these rules will make a mistake and instead go wrong, and worse yet, when they do so they will also mistakenly think that they are

[77] Thanks to an anonymous referee for pressing me on the distinction between conclusions about E type evidence generally and conclusions about any specific piece of E-type evidence.

getting it right. But this regrettable behavior is no mark against those who are instead epistemically excellent.

My goal now will be to dispel the appeal of this line of response. One strategy for doing so focuses on the last concession: we might think that theories of epistemic rationality shouldn’t be allowed to so cavalierly wash their hands of those who try, but fail, to live up to their standards. That will not be my focus. Rather, I want to attack BT views on their most favorable grounds. So I will, in pursuing my response, focus exclusively on the best case for the BT views: the case where Misty really has evaluated every piece of E-type evidence she’s received rationally, and where she in each case has correctly ascertained that she has done so. Even in this case of perfect performance, I think the BT view significantly underestimates the difficulties of combining a recognition of one’s general fallibility in handling E-type evidence with a policy of uncritically responding to it in each particular instance. To draw out these difficulties, I start by considering what life would be like for someone who accepted my general conclusions about E-type evidence, but then tried to hold on to their non-conciliatory view in this way. My first claim will be that such a person is committed to pervasive self-binding; illustrating that will be the task we now turn to.

4.11 BT VIEWS LEAD TO SELF-BINDING

My general strategy for illustrating the self-binding BT views require will be to return to our central example, of Misty taking A Frustrating Final, and to consider a series of bets that we might offer her on whether she or Ash has gotten the right of things. As we have established, earlier in the process she anticipates that responding to the E-type evidence she is bound to receive will degrade her accuracy. It follows that she thereby expects decisions she makes on the basis of those later credences to do worse. Since she expects decisions based on her later credences to do worse, she should now be willing to do her best to preempt them. So, for instance, she should be willing to pay money in order to constrain or eliminate choices she might later be offered. As such, she will be interested in binding her future self to her current will.

To illustrate: suppose Misty has thought things through and come to adopt the BT-based response I have outlined about. She accepts that in general trying to respond to E-type evidence will lower her accuracy, but nonetheless holds a BT type view and thinks that in each instance, what she ought to do is respond to E *correctly*. Now also suppose that I am lurking about, waiting to make trouble. I tell Misty "I am going to watch you and Ash leave the lecture hall after your final examination. I will wait for you to compare your answers, and then—if I see the look of mutual dismay on your faces that indicates you disagreed—I will spring from the bushes! And I will offer to you a deal. I will sell you a ticket that pays out a dollar if, once the grades come in, it turns out you were the one who was wrong on that final. I will offer it to you at the low, low price of 50 - N/2 cents."[78]

Misty knows she will reject that offer. She knows that in the situation described, if it occurs, she will think that she is 50 + N% likely to be right. Therefore, at the time that she is offered this bet she will think the expected value of the ticket is only 50 - N cents, and so decline the opportunity to purchase it.

But: that is then and this is now. Currently, Misty still thinks it's 50 / 50 whether she'll be the wrong one should a disagreement arise, and so thinks that the ticket has an expected value of 50 cents. As such, she thinks that although she unfortunately *won't* be willing to buy it, if she *were*

[78] One might object: money is not continuously graded, so if N is small enough there's no guarantee that there will be appropriate cent values to make the argument go through. It would be curious, however, to argue that non-conciliatory views are defensible because the money we contingently use is coarsely graded.

to do so she would net an expected gain of N/2 cents. Misty thinks that if I do what I say, the result will be that she misses a good deal.

But if that's really what she thinks, it makes sense for her to act to try to secure that good deal. For instance, Misty could counter-propose to me the following, slightly more complicated bet: first, she pays me any amount up to M% of N/2 cents. Then, if she and Ash wind up disagreeing, she must pay me an additional 50 - N/2 cents. Thereafter, when the grades come in, if it turns out that Misty was the one that was wrong, I pay her a dollar. Why should she be willing to make this counter-proposal? Because on her current understanding of the situation, it nets an expected profit of whatever difference remains between her initial payment and M% of N/2 cents.

That's all a mouthful. But the informal description of what Misty is doing by way of this counterproposal is straightforward. Misty is paying me some money now in exchange for my later forcing her to take that bet, the bet she otherwise would decline. Since the amount of money she's paying me is smaller than the expected profit from that bet, she should think that even minus the payment she'll still net a gain. Current Misty is not willing to leave things up to future Misty; she wants to lock in her choice now, because she knows what future Misty *will* believe, and she currently takes it to be a worse basis for decision making than her current credences.

4.12 SELF-BINDING AND CONCILIATION

If Misty accepts my characterization of E-type evidence, yet nonetheless holds on to her BT-view, she will thereby be committed to self-binding. This is not yet a reductio of her position. Those with a fondness for BT views may take the argument thus far not to refute them, but merely to be an interesting discovery: it turns out, because BT views are true, that cases like A Frustrating Final are cases where self-binding is rationally required. Call this the BTS view, as it is the BT view as supplemented by self-binding.

We can illustrate this thought with an example from Roger White, who himself appears to hold a BTS-style view (White 2010: 601-604): it may make sense at the beginning of the night, given how much you anticipate drinking, to give your keys to your friend and tell them not to let you drive home. You may do this because you think that later in the night you will be drunk, and there's a significant risk that you may misevaluate your state and consequently try to drive. But suppose, contra your expectation, the night passes tamely and at the end of it you are rather sober. You may not only *be* sober, but *know* that you are. If so, then you may rationally seek to do what you earlier rationally tried to restrict yourself from doing, e.g., get your keys from your friend and drive home.

The BTS view assimilates cases of peer disagreement to this model. It holds: *in general* E-type evidence is such that responding to it reduces expected accuracy. And for that reason it makes sense, ahead of time, to take measures to prevent yourself from making decisions on its basis. But once you get any particular piece of E-type evidence, you may not only respond rationally to it, but you may furthermore see that you have done so. And once you see that, you may rationally go on to try to make the very decisions you earlier were rationally trying to restrict yourself from. There is a sort of practical friction here, but, the epistemic asymmetry underlying the practical friction is well motivated and so, the thought goes, there's nothing wrong with it.

Now, before going on to give my argument against BTS views, I want to remark on why, even if they are in the end correct, they still preserve a significant conciliatory spirit. Much of the interest that drives the peer disagreement debate is practical. As such, if it turned out that we ought to make sure we conciliate—not because it is rational in the moment, because it isn't, but rather because it is rational to make ourselves into the sort of creatures who will be constrained to conciliate in the moment—it would still be true that we should try our very best to see to it that we go on to conciliate. And learning that we ought to make ourselves into conciliators would have deep practical consequences.[79]

So I think it is important to recognize that BTS-licensed conciliatory self-binding, when genuinely pursued, may result in a surprisingly robust conciliatory program. In pointing this out, though, I do not mean to suggest that BTS views will wind up having identical practical upshot to conciliatory views: exactly how close they wind up depends on the answers to questions I won't enter into here.

I should note, however, one thing that I *do* assume about the self-binding BTS views endorse. Namely, I assume that it is not adequate to bind oneself merely by making a conscious decision to set a policy. I take it that this is what gives the BTS view its distinctive non-conciliatory flavor. If all it took to bind ourselves to a policy was to select it by some conscious act, then the difference between a program BTS-licensed conciliatory self-binding and conciliation proper would risk becoming slim indeed. In any case, I am going to assume that the sort of binding a BTS-er imagines as effective is e.g. the sort relayed in White's story where one commissions a friend to act as controlling supervisor of one's keys, or in my earlier discussion, where one uses pre-betting to effectively take decisions out of one's future hands.

[79] Compare with Newcomb cases: if we think two-boxing is rational but nonetheless we prudentially ought to see to it that we become one-boxers, this conclusion is not very practically compelling. After all, we do not ever expect to face a Newcomb case.

4.13 BTS VIEWS DEPRIVE US OF THE BENEFITS OF OUR FACULTIES

Why not hold a BTS view? In this section I argue that BTS views construe our epistemic predicament as tragic. I hold, however, that we have no compelling reason to accept this tragedy. So we ought to reject it, and BTS views along with it.

What is the alleged tragedy? If BTS views are true, then there are cases in which Misty thinks her powers of judgment are a valuable indicator of the truth about a question she is deeply interested in, yet nonetheless she will rationally do her best to avoid deploying them. She will choose instead to carry on in ignorance. Rationality will thereby require her to lock herself out of the benefits of her own faculties.

The case I'll use to demonstrate is similar to the earlier Frustrating Final insofar as it involves Misty thinking ahead of time about a future disagreement; it is more complicated, however, insofar as we consider a situation in which more than one peer is at work. So imagine that we have *A Crowd of Experts:* Misty holds the BTS view. She is also an expert mathematician, and she and her eight friends work in an obscure subfield. Recently, she has become aware that they are abuzz over a new result; four of them think that the proof of this result is valid, four of them are convinced that it relies on a subtle equivocation. Misty doesn't yet know what the result is or how the putative proof goes, but she thinks of herself and her friends as independent and highly reliable judges; given that they are evenly divided in this early stage of investigation, she takes it to be 50 / 50 whether the proof is valid. Misty, anticipating that she will soon be asked to investigate the proof herself, considers what will happen when she does. She anticipates that she will find it either valid or find it invalid. Conceiving of this as a future exercise of her mathematical judgment, in the abstract, she thinks of it as on a par with the judgements of any of the other experts who have already evaluated the proof. So, she anticipates that there will be a 5-4 split of the relevant experts, and she thinks that, under those conditions, the odds rise to 65 / 35 in favor of whichever option she judges, given that it will *ipso facto* command the slight majority. She also thinks that she will, after forming her judgment, not merely respond to the fact that she so judged, but she will also respond to the richer E-type evidence that will then become available to her. She anticipates that she will, in so doing, become predictably convinced that it is 85 / 15 in favor of the option she judges (that is to say, the predictable extra weight she will accrue in the course of applying her BTS view will amount to another 20% confidence—or, in my earlier terms, N is .2).[80]

Now consider the following development, where *Malevolent Aliens Force Practical Consequences:* Misty is abducted by powerful, malevolent, and mathematically sophisticated aliens. They inform her that they have long ago considered the mathematical result being discussed by her colleagues, along with the attendant line of proof, and they know with perfect certainty whether it holds. They offer to show her the purported proof her colleagues are puzzling over and give her a chance to think it through. They also inform her that, in a week, they will ask her how likely she thinks it is that the proof is valid. They will then take her answer, measure its inaccuracy using the Brier score, then multiply the result by 10,000 and murder that many humans. She has every reason to believe this is true.

Finally, suppose there are *No Restraints:* Stranded as she is, Misty has no available means to bind her future self to any current decisions. There's no time to instill habits; the aliens are not susceptible to elaborate insurance bets; she cannot pre-select answers, etc. etc. Her only choice is over whether or not she wants to see the purported proof before she reports on whether it is sound.

[80] The situation so described is not intended to be a general description of mathematical epistemology, e.g. there is no reason to suppose that it will always be true that one ought to suppose, when mathematicians evenly divide over the validity of an early proof, that it is thereby 50 / 50 whether it is valid, nor is there any special connection between a 5-4 split and the particular 65 / 35 or 85 / 15 odds assigned. The numbers in this story are not supposed to be deduced from one another; they are just values which we have no reason to think couldn't arise.

In the case as described, Misty ought to decline to see the purported proof. Given that she holds a BTS view, she anticipates that she will predictably accrue extra weight to her assessment. And in this case, that extra weight is such that her expected accuracy afterward will be worse than her current 50 / 50 guess. As such, in declining to see the proof she minimizes the expected death count.

Again we can describe, on the basis of this information, what Misty anticipates will happen if she is shown the purported proof. She currently takes it to be 50 / 50 whether the proof is valid. Given that she is currently indifferent about the validity of the proof, something which she takes her judgment to be an imperfect indicator for, she is also indifferent about whether she will judge it valid or invalid. So the first branch in the possibility space is the 50 / 50 chance she assigns to her judging the proof valid. From there, there is another branch representing how likely she takes it to be that the proof *is* valid, given her judgments; the odds off these branches are 65 / 35 in each case that she is getting it right, as that's what she takes the 5-4 split of experts that will then exist to indicate. But regardless of how this goes, in all cases she expects to form a .85 credence that whichever option she has judged is correct; this is the additional extra weight that she will accrue to her view in attempting to respond to the richer evidence available to her when she sees the actual proof.

If we first multiply out along these branches, and then we apply Brier scoring, we can arrive at the conclusion that the expected inaccuracy of Misty's .85 credence she knows she will acquire after being shown the proof is .26732. Misty's current inaccuracy, given her 50 / 50 credence, is Brier scored as .25. So Misty's expectation of her inaccuracy upon seeing the proof is .1732 worse, corresponding to a worsening of the expected death toll by 17,320 people.

I have constructed this example with particular numbers, and I have used a particular scoring rule—the Brier score—but the structure is general. If BTS views are true, we should expect such cases to arise. What generates the accuracy deficit is the predictable extra weight Misty expects to accrue, and that there is such predictable extra weight is just the same core feature of non-conciliatory views we already explored in the context of A Frustrating Final.

Rather, what is interesting and new about the way we've filled in the case is that we've set the value of 'N'—the predictable extra weight Misty will accrue—relatively high. The extra weight here is pretty heavy.[81] And that is the important difference with A Frustrating Final. In A Frustrating Final, we showed that the evidence E non-conciliatory theorists appeal to is such that responding to it lowers Misty's expected accuracy, and hence, if she could choose ahead of time to self-bind to avoid responding to it she would. However, in that case the E-type evidence she receives comes packaged with some other evidence—namely, the facts about distribution of expert judgment that she acquires over the course of thinking through the problem and then meeting up with Ash. So for all we have said, it may well be that even though the E-type evidence alone has a negative effect on her expected accuracy, that's outweighed in the final balance by the rest.

By contrast, here we have fixed the pernicious effects of the E-type evidence to be significant enough that Misty not only anticipates that responding to it will lower her expected accuracy, but that this bad effect will be large enough to outweigh the other benefits Misty gets from working through the problem and adding her expert judgment to the stock of those already existing. So Misty should not want to work through the problem and add her expert judgment to

[81] Perhaps this is illicit? I think not. In cases of deductive reasoning, like the one we are presently considering, numbers like these make sense; deductive reasoning will often be such that one's individual judgment considered as such is a weak indicator, but where the first-order considerations are very powerful—deductive arguments are, after all, as strong as first-order considerations get. So the difference between the conciliatory response and the response that is justified by the first order considerations should be very large, and so N should be able to range quite high.

the stock of those already existing: she should prefer to avoid thinking about it altogether, and just answer the aliens on the basis of what she already knows about how the *other* experts have answered.

On reflection, however, we should see this as a perverse result. After all, Misty is a mathematical expert. Everyone, herself included, takes her judgment of the purported proof to be a genuine indicator with respect to its validity. And yet, when it comes time to make this important decision, in which many lives depend on the validity of the purported proof, she'd rather not draw on her own judgment. She'd rather make her decision *only* on the basis of the eight other expert mathematicians' judgments.

She does not think those other experts judgments are any better than hers, and when she thinks about expert judgement in the abstract, more judgments is always better. If she could have chosen for another expert to have seen the proof, judged it, and then had their opinions be reported to her, then she certainly would have. But despite being just such an expert herself, she cannot make use of her own judgment.

And this, I take it, is the sense in which she is deprived of the use of her own faculties. Because she subscribes to the BTS view, Misty must leave her powerful and relevant mathematical abilities fallow, even as she is confronted with precisely the sort of problem to which they are so well-suited to contribute.

4.14 BTS VIEWS' TRAGEDY IS UNDERMOTIVATED

As an acolyte of the BTS view, Misty will decline to see the proof. But what if she winds up seeing it anyway? Suppose the aliens show it to her against her wishes. If so, then she will evaluate it and come to be .85 confident in its (in)validity—as laid out in the previous diagram. In this section I want to suggest that there is something very odd about this combination.

Return to an earlier example: I have left some possessions stranded at the house of my ex-boyfriend. I want to get them, but I also know that he is an expert at manipulating me and so do not want to give him the chance to speak to me.

How should I think about his hypothetical speech in advance? I should think that I will almost certainly not respond to it rationally. Even though I am highly confident that nothing he says will present a compelling reason to believe him, I still think there's a significant chance I will believe him anyway. So I anticipate that my response to hearing him out will be irrational in a disastrous way.

At the same time, I should also think that his speech will consist in perfectly good evidence. After all, the content of his speech will likely give a picture of his state of mind and show with greater clarity the emotional tactics he uses to try to control me. And, though it is long odds, there is even the remote chance that his speech would contain genuine evidence that I ought to do and believe as he wants—it is at least *conceivable* that there could be such a reason, and that he could explain it to me.

Given this set-up, ask: should I want my beliefs to be informed by responding to my ex's speech? There is a tension here. On the one hand, it's true that the speech is in some sense good evidence. On the other, though, I anticipate that my response to it will nonetheless be very bad. Here are two coherent ways to resolve this tension:

First, we might emphasize that the speech is good evidence. As such, we may say: of course I should want my beliefs to be informed by responding to it. So I should make sure that I get the chance to hear it; I should seek him out and listen attentively. Once I've listened to his speech, I should then carefully think it over and apportion my beliefs as closely to this new evidence as I possibly can. I should always want my beliefs to be informed by good evidence. I find this answer very difficult to believe in light of the case—as I've already made clear in section 4.2—but nonetheless it at least presents a consistent view.

Second, we might emphasize that I anticipate my response being bad; I am very confident that I will be convinced to believe something terrible. As such, we might say: I ought not want my beliefs to be informed by responding to this evidence, because that will likely make them terrible. So I should make sure that I do *not* get a chance to hear my ex's speech. And supposing that somehow, against my will, I do manage to hear it, I should make sure that I don't let it influence me. If he manages to accost me, I should ignore him. I should ignore my own impulse to believe him, and so on—at least best as I can. I think this, as I have again made clear, is the better answer.

But regardless of which is better, both answers are consistent in the following way: they give a unified answer to *whether I should want to hear the speech* and *whether I should listen to it if I do.* They answer, naturally, in terms of whether I should want my beliefs to be informed by my ex's speech. If yes then, as in the first response, I should both seek out and listen to it. If no then, as in the second response, I should both avoid and do my best to ignore it.

The BTS view, by contrast, splits its answers. It tells Misty: make sure you don't see the proof if you can help it. But once you do, you should carefully evaluate it and apportion your beliefs as appropriate. But this mix looks odd when we hold it up as against the purer alternatives. Who would say: you ought to avoid your ex, surely, but if he manages to catch and corner you then you should carefully evaluate what he has to say?

After all, it seems that the very same reasons I have for not wanting to run into my ex count just as much in favor of trying to ignore him once I do. And given that the very same reasons are operative for each—reasons relating to our anticipated rational failures—we can even think, if we like, of our ability to ignore as an internal counterpart of external avoidance strategies. One way we can deal with the fallibility of our rational capacities is by avoiding the sort of evidence that we anticipate will lead us astray altogether, as, for instance, when we try to avoid coming into contact with manipulative exes. But another way is by allowing ourselves to encounter the evidence, but then exercising higher-level control to stymie our responses to it, as we do when we encounter those exes but do our best to muffle our impulses to believe them. The BTS view fails to notice the continuities in both purpose and effect of these strategies.

The crowd of experts case is designed to highlight the arbitrary and inadequate nature of purely external management. In that case, Misty's external management strategies are limited to two options: either she can see the purported proof, or she cannot. If she sees the purported proof, she will get both the evidence consisting in her expert judgment of its validity, and also the richer E which includes whatever counts as the first-order evidence in a case of mathematical proof. She can't get one without the other, because she can't generate her expert judgment of the purported proof's validity without actually wading through the relevant E. So she has no way to use external management to achieve the result that she anticipates would actually be best: getting an additional expert judgment on the soundness of the proof without accruing all the extra weight from that E.

If she allows herself to practice internal management, though, she can. She can get the benefit of her expert judgment without being swayed by the co-occurring E-type evidence: she can do so just by using that evidence to form her judgment, but then ignoring it thereafter.

As far as I can tell, BTS views have precisely one objection to this internal management: when Misty ignores the full force of her evidence, she misses out on the benefits she could get from it. But it is worth remembering that, when we look at Misty's passage through the whole of the case, *regardless* of which view she follows she will lose out on the full potential benefits of incorporating E into her beliefs. After all, on BTS views, when things go according to plan she will never even get E in the first place. Rather, she will decline the aliens' offer to see the proof. BTS views need it to be the case that there is a distinction between two *ways* of excluding E from being factored into her final beliefs, and they need this distinction to be significant enough to justify pursuing purely external management even when internal management would allow for better results. This is hard to sustain when: theoretically, external and internal management share common purposes and effects, and intuitively, when we look to cases like that of the manipulative ex, the strategies seem to go hand in hand.

Thus concludes my brief against BTS views. On such views, Misty's situation is epistemically tragic insofar as she cannot make good use of her own expertise. This tragedy rests on a distinction between internal and external tactics for managing one's beliefs. I say we ought to reject both the distinction and the tragedy. Misty ought to get use out of her faculties, and she ought to do so by managing their deployment in the way the conciliationist suggests.

I spent this time going over BTS views because they represent, I think, the best chance for conceding that E-type evidence is deleterious to expected accuracy, yet blocking the further inference to conciliation. Having argued that BTS views fail, I thereby take myself to be entitled to reinstate that inference. E-type evidence is deleterious to expected accuracy, and so we ought to ignore it. Once we do, all that's left to respond to is the mere fact of disagreement—and once we limit ourselves to that, we are conciliationists.

4.15 REFLECTION AND CONDITIONALIZATION

We have traversed a great deal of ground. It is worth taking a moment to situate my argument in relation to some other, perhaps more familiar, epistemic claims and arguments.

I have argued that non-conciliatory views take what they foresee to be bad epistemic deals, and that these deals look bad as a result of a disagreement between Misty's epistemic points of view at two different times. Right now Misty thinks there's a 50 / 50 chance that if she and Ash disagree tomorrow she will be right, but at the same time she thinks that if she and Ash disagree tomorrow she *will, at that time,* believe that there's a more than 50 / 50 chance she's right. Since she does not now accept those odds that she knows she will later accept, she perforce thinks the later odds are less accurate. This generates the funny behavior.

One might think, then, that what's needed is a principle guaranteeing regularity between present and future beliefs. A commonly discussed family of such principles are *reflection* principles.[82] For instance, one such principle might be: if you have good reason to think that tomorrow you will believe *p*, believe *p* now. After all, suppose I tell you "tomorrow you will think it's raining." Then isn't that a good reason to think it will rain tomorrow? After all, usually you think things because they're so. So, we might ask, can the force of the argument given in this paper be captured by such a reflection principle?

The answer is no. All plausible versions of reflection principles include exemptions for, among other things, irrationality. Suppose I tell you "tomorrow you will be captured and brainwashed into thinking Obama is a lizard person." This is not a good reason to now think Obama

[82] For an argument against steadfast views that does go by way of reflection, see (Setiya 2012). For an original locus of discussion on reflection principles see (van Fraassen 1995).

is a lizard person.[83] As such, reflection principles will only demand matching between present and future when one has good reason to think that one's future beliefs will be rational. But, at least on the total evidence and right reasons views, this will not be so in cases of peer disagreement: one will think, rather, that it's only 50 / 50 whether one's future attitude will be rational. So even if some reflection principle is true, it will not be able to tell against those views. Yet the argument in this paper does tell against those views. Therefore, the argument in this paper cannot be captured by some form of reflection principle.

I have put my critique in terms of expected accuracy. Such a framing might lead one to wonder: there already exist arguments in the literature which purport to show, with mathematical clarity that expected accuracy is uniquely maximized by holding a probabilistically coherent set of credences and then updating them by conditionalization.[84] Together, these two claims—probabilism and conditionalization—constitute the traditional Bayesian package, and so, it might be thought, these arguments show that anyone concerned with maximizing their expected accuracy ought to become some stripe of Bayesian. But if this is so, I am in trouble, as investigation reveals that traditional Bayesianism is actually inconsistent with conciliation.[85] If Bayesians have a mathematically sound monopoly on expected accuracy, then in trying to use expected accuracy to argue for conciliation I must have gone terribly wrong somewhere.

The answer is that those arguments do not actually show that anyone concerned with maximizing expected accuracy ought to become some stripe of Bayesian. Take conditionalization: such arguments may show that *actually* updating by conditionalization maximizes expected

[83] See (Christensen 1991) for a compelling and detailed presentation of this point.

[84] For the argument for probabilistic consistency, see (Joyce 1998); for the argument for Conditionalization see (Greaves and Wallace 2006).

[85] c.f. (White 2009) on the incompatibility of the 'calibration rule' with Bayesian epistemology.

accuracy. However, they do not show that *trying* to update by conditionalization maximizes expected accuracy. And in many cases we have decisive evidence that if we *try* to update by conditionalization, what we will in fact do will be something else entirely.

We can again frame this in terms of the examples from the beginning of this paper. Suppose I listen to the testimony of my ex-boyfriend, or view the races and genders of all the applicants auditioning to my orchestra: I have strong evidence that if I do so, I will not react by conditionalizing on that new evidence. Rather, I have strong evidence that I will react by being convinced to return to my ex, or by forming more negative evaluations of the minority applicants' talents—and those are beliefs I now anticipate to be less accurate than my current ones.

I take it that the fact that *succeeding* at conditionalization maximizes expected accuracy is of great epistemic interest. Still, there are lots of things it would be lovely to succeed at, but that it's nonetheless best not to try—because one will likely fail, and the costs of failure will be significant. For instance, even if actually doing a backflip would impress everyone in the room, I ought not try. I take it that when considering what we should believe, just as when considering what we should do, we ought to take account of evidence that we will not succeed at doing what we try. In these cases, I have excellent evidence both that I will not succeed at conditionalization and that the results will be bad. So I should avoid trying to conditionalize on that evidence.[86]

[86] Here I am in substantial agreement with (Schoenfield forthcominga). I take it that both the rules it would be best to follow and the rules that it would be best to try to follow will each play substantial roles in our total epistemic theory—and, furthermore, that it is an attention to the latter which motivate conciliation. Thus I agree that conciliation ought to be grounded in a 'trying' account; I take it such an account will govern the prescriptive 'should' of 'should believe.'

4.16 CONCLUSION

When Misty finds that Ash disagrees with her over the answer to their frustrating final, what should she believe? Should she take things to be 50-50, as she antecedently expected they would be in light of his disagreement, or does she need to adjust that expectation in light of further evidential features of the case? Conciliationists and their opponents have clashed over the proper characterization of the evidence at Misty's disposal, under the assumption that answering the question of what her evidence supports would straightforwardly answer the question of what she ought to believe. If her evidence is exhausted by the mere fact of disagreement then conciliation stands; if her evidence outstrips the mere fact of disagreement then conciliation falls.

Evidence is being afforded a central role in the debate, but it is worth taking a step back and asking why it is that we care about evidence in the first place. I have framed the positive role of evidence in terms of a process of inquiry we value for its tendency to lead us toward truth and away from error. But if this is really what is valuable about getting and responding to evidence, then conciliatory answers to what Misty ought to believe can survive even quite a non-conciliatory construal of the evidence available in disagreement cases. I have argued: construing such evidence as having anti-conciliatory force at the same time make the evidence such that Misty expects trying to respond to it to lead her into error. So, from her perspective, any such anti-conciliatory evidence thereby lacks the truth-conducivity that makes evidence worth paying attention to in the first place.

And that, then, is my ultimate conclusion. If there is any anti-conciliatory evidence, then it is highly unusual in precisely such respects that we should not want to respond to it. So we shouldn't respond to it and instead we should conciliate.

BIBLIOGRAPHY

Berker, Selim (2008). Luminosity Regained. *Philosophers' Imprint* 8 (2):1-22.

Christensen, David (1991). Clever Bookies and Coherent Beliefs. *Philosophical Review* 100 (2):229-247.

Christensen, David (2007). Epistemology of Disagreement: the Good News. *Philosophical Review* 116 (2) 187-217.

Christensen, David (2009). Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass* 4 (5):756-767.

Christensen, David (2010). Higher-Order Evidence. *Philosophy and Phenomenological Research* 81 (1):185-215.

Christensen, David (2011). Disagreement, Question-Begging, and Epistemic Self-Criticism. *Philosophers' Imprint* 11 (6).

Christensen, David (forthcoming).Disagreement, Drugs, etc.: from Accuracy to Akrasia. *Episteme*.

Cohen, Stewart (2010a). Bootstrapping, Defeasible Reasoning, and A Priori Justification. *Philosophical Perspectives* 24 (1):141-159.

Cohen, Stewart (2010b). Luminosity, Reliability, and the Sorites. *Philosophy and Phenomenological Research* 81 (3):718-730.

Elga, Adam (2007). Reflection and Disagreement. *Noûs* 41 (3):478–502.

Elga Adam (2010a). How to Disagree About How to Disagree. In Ted Warfield & Richard Feldman (eds.), *Disagreement*. Oxford University Press. 176-186.

Elga, Adam (2010b). Subjective Probabilities Should Be Sharp. *Philosophers' Imprint.* 10 (5).

Elga, Adam (2013). The Puzzle of the Unmarked Clock and the New Rational Reflection Principle. *Philosophical Studies* 164 (1):127-139.

Enoch, David (2010). Not Just a Truthometer: Taking Oneself Seriously (But Not Too Seriously) in Cases of Peer Disagreement. *Mind* 119 (476):953 - 997.

Fumerton, Richard (2009). Luminous Enough For a Cognitive Home. *Philosophical Studies* 142 (1):67-76.

Gibbard, Allan (2003). *Thinking How to Live.* Harvard University Press.

Gibbard, Allan (2007). Rational Credence and the Value of Truth. In Tamar Szabo Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology: Volume 2*. Oxford University Press. 143-164.

Goldin, Claudia and Rouse, Cecilia (2000). "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians. *American Economic Review* 90 (4):715-741.

Greaves, Hilary & Wallace, David (2006). Justifying Conditionalization: Conditionalizing Maximizes Expected Epistemic Utility. *Mind* 115 (459):607-632.

Hawthorne, John & Srinivasan, Amia (2013). Disagreement Without Transparency: Some Bleak Thoughts. In David Christensen & Jennifer Lackey (eds.), *The Epistemology of Disagreement: New Essays*. Oxford University Press. 9-30.

Hawthorne, John & Stanley, Jason (2008). Knowledge and Action. *Journal of Philosophy* 105 (10):571-590.

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

Joyce, James M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science* 65 (4):575-603.

Kelly, Thomas (2005). The Epistemic Significance of Disagreement. In John Hawthorne & Tamar Gendler (eds.), *Oxford Studies in Epistemology, Volume 1*. Oxford University Press. 167-196.

Kelly, Thomas (2010). Peer Disagreement and Higher Order Evidence. In Alvin I. Goldman & Dennis Whitcomb (eds.), *Social Epistemology: Essential Readings*. Oxford University Press. 183-217.

Kelly, Thomas (2013). Disagreement and the Burdens of Judgment. In David Phiroze Christensen & Jennifer Lackey (eds.), *The Epistemology of Disagreement: New Essays*. Oxford University Press. 31-53.

Lackey, Jennifer (2008a). A Justificationist View of Disagreement's Epistemic Significance. In Alan Millar Adrian Haddock & Duncan Pritchard (eds.), *Social Epistemology*. Oxford University Press. 145-154.

Lackey, Jennifer (2008b). What Should We Do When We Disagree? In Tamar Szabó Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology: Volume 3*. Oxford University Press. 274-293.

Lam, Barry (2011). On the Rationality of Belief-Invariance in Light of Peer Disagreement. *Philosophical Review* 120 (2):207-245

Lam, Barry (2013). Calibrated Probabilities and the Epistemology of Disagreement. *Synthese* 190 (6):1079-1098.

Lasonen-Aarnio, Maria (2011). Unreasonable Knowledge. *Philosophical Perspectives*. 24 (1): 1-21.

Lasonen-Aarnio, Maria (2014). Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research* 88 (2):314-345.

Lord, Errol (2013). From Independence to Conciliationism: An Obituary. *Australasian Journal of Philosophy* (2):1-13.

Setiya, Kieran (2012). *Knowing Right From Wrong*. Oxford University Press.

Schafer, Karl (2014). Doxastic Planning and Epistemic Internalism. *Synthese* 191 (12):2571-2591.

Schafer, Karl (forthcoming). "Self-Trust and Rational Symmetry." *Philosophy and Phenomenological Research*.

Schechter, Joshua (2013). Rational Self-Doubt and the Failure of Closure. *Philosophical Studies* 163 (2):428-452.

Schoenfield, Miriam (2012). Chilling Out on Epistemic Rationality. *Philosophical Studies* 158 (2):197-219.

Schoenfield, Miriam (2014). A Dilemma for Calibrationism. *Philosophy and Phenomenological Research* 89 (2): 425-455.

Schoenfield, Miriam (forthcoming a). Bridging Rationality and Accuracy. *The Journal of Philosophy.*

Schoenfield, Miriam (forthcoming b). Internalism Without Luminosity. *Philosophical Issues.*

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

Smithies, Declan (2012). Mentalism and Epistemic Transparency. *Australasian Journal of Philosophy* 90 (4):723-741.

Srinivasan, Amia (2013). Are We Luminous? *Philosophy and Phenomenological Research* 90 (1): 294-319.

Titlebaum, Michael (2015). "Rationality's Fixed Point (Or: In Defense of Right Reason)." In John Hawthorne & Tamar Gendler (eds.), *Oxford Studies in Epistemology: Volume 5*. Oxford University Press. 253-294.

van Fraassen, Bas C. (1995). Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies* 77 (1):7-37.

Weatherson, Brian. Do Judgements Screen Evidence? Manuscript.

Weatherson, Brian (2013). Disagreements, Philosophical and Otherwise. In Jennifer Lackey & David Christensen (eds.), *The Epistemology of Disagreement: New Essays*. Oxford University Press. 54-76.

Williamson, Timothy (2000). *Knowledge and its Limits.* Oxford University Press.

Williamson, Timothy (2005). Replies to Commentators. *Philosophy and Phenomenological Research* 70 (2):468-491.

Williamson, Timothy (forthcoming). Very Improbable Knowing. In T. Dougherty (ed.), *Evidentialism and its Discontents.* Oxford University Press.

Willenken, Tim (2011). Moorean Responses to Skepticism: a Defense. *Philosophical Studies* 154 (1):1-25.

van Wietmarschen, Han (2013). “Peer Disagreement, Evidence, and Well-Groundedness.” *Philosophical Review* 122 (3):395-425.

Wedgwood, Ralph (2010). The Moral Evil Demons. In Richard Feldman & Ted Warfield (eds.), *Disagreement*. Oxford University Press. 216-246.

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

White, Roger (2010). You Just Believe That Because…. *Philosophical Perspectives* 24 (1):573-615.