AGAINST RIGHT REASON

Robert Steel

Abstract: I argue against 'right reason' style accounts of how we should manage our beliefs in the face of higher-order evidence. I start from the observation that such views seem to have bad practical consequences when we imagine someone acting on them. I then catalogs ways that Williamson, Weatherson, and Lasonen-Aarnio have tried to block objections based on these consequences; I argue all fail. I then move on to offer my own theoretical picture of a rational 'should believe,' and show that, if such a picture is right, it can neatly explain why right reason isn't. I close by arguing that the extent to which anti-luminosity arguments motivate right reason has been overstated; the positive picture developed here, despite rejecting right reason, is nonetheless consistent with luminosity failures.

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:

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. 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:

> *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,"[1] as well as "first-order" and "higher-order."[2] 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.'[3] For my purposes here, it's enough to point them out by ostension: actual justification encompasses the

---

[1] Willenken (2011)

[2] e.g., Christensen (2010), and, seminally, Kelly (2005).

[3] I take the terminology of 'calibration' and 'calibrationism' from Schoenfield (2014). I choose the term "actual justification" because, in paradigm cases like that of the worried parent, there is often a temporal gap where one becomes uncertain whether the features that one initially responded to did, in the context of one's initial response, *actually justify* the attitude one formed. But I intend nothing of substance to rest on this terminological choice.

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 ascertain it.[4]

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.[5]

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.*[6] And, of course, intermediate views are also possible. What a person should believe might be some non-trivial function of *both* the actual justification and calibration. Elsewhere, I refer to that family of views as *interactionism.*[7]

For full disclosure,I am sympathetic toward calibrationism. But 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

---

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

[5] The term "right reasons" was coined by Elga (2007) to describe Kelly (2005)'s seminal articulation of a view of this sort. Both Elga and Kelly then put the dispute in terms of evidence, and as noted in the main I intend a more agnostic framing on that point. Nonetheless, Kelly (2005) is a good early example of the sort of view under attack here.

[6] 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.

[7] In my "Against Interactionism," currently in draft.

argument for calibrationism, though, by eliminating one of the main structural alternatives.

Here's the plan: in section 2, I introduce the right reason view and survey some of its motivations. In section 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 practical results in a wide range of problem cases. Then, in section 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 5 and 6 I consider some existing proposals along these lines, but argue that they fail to meet the criteria I've identified. In section 7 I outline what I take to be a better, right reason-unfriendly, explanation of the problem cases. Finally, in section 8, I consider and respond to a line of objection based on Williamson's anti-luminosity arguments before concluding.

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: justified beliefs 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.[8] 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.[9]

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 transmitting well-foundedness from a group of premises to a conclusion 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 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

[8] 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. So although 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.

[9] Lasonen-Aarnio (2010) 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 ftn. 42 for further discussion of this argument.

be drawn to think the same of actual justification generally.[10]

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.[11]

And on some views, what goes there for the mathematical example will go for all inductive 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.[12] So, if one finds these views attractive, then one may be

[10] 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 (2014) 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.

[11] Weatherson (ms.) makes a point like this: he 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.

[12] 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).

drawn to the right reason view.[13]

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 probabilistically coherent, such that the evidence always supports confidence in a logical consequent at least as high as it supports those antecedents that entail it. And I think that this remains true of *evidential support* regardless of how long or difficult it would be to actually offer a proof of that consequence relation. I also believe that relations of evidential support will turn out on the best understanding to be both *a priori* and empirically indefeasible.[14] 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.

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

[13] 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.

[14] To emphasize, my aim here is not to offer any serious defense of these claims about evidential support. I raise them just to highlight the potential to co-opt some of the motivations for right reason by introducing further distinctions between epistemic concepts.

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:[15]

Successful Reasoner: 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.

Overwhelming Adverse Calibration: before she starts construction, her doctor confronts her. He claims to have discovered from recent 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.[16]

Finally, we add some consequences to make sure the outcome matters.

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?

[15] These cases can be multiplied endlessly; this one is a close variant of Hawthorne & Srinivasan (2014).
[16] 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.

Now ask: 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 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.[17] 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

[17] see e.g. Christensen (2010) and Schechter (2013).

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.

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.[18] 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.[19] 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

[18] Specifically: the claim of *default* matching, and the request for explanations of disconnect, is thoroughly neutral with respect to pragmatic encroachment. What follows neither presupposes nor denies its possibility.
[19] Williamson (2005, p.479–483).

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. Nonetheless, 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.

On the view being developed, 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?[20] Williamson's discussion has the right potential shape: it makes a distinction that purports to show

[20] Williamson endorses the right reason view for at least one class of cases: closure cases, as noted in ftn. 10. 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

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 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

---

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.

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.[21]

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

In the next two sections I 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

[21] 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 p.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 8

concept that seems to be a better match for the epistemic 'should' we are interested in.

5. Blocking the Simple Argument: Incongruous Higher-Order Beliefs; Practical Bridge Principles

I begin with Weatherson's proposal for handling the simple argument. 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[22]—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-

[22] His discussion in 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 section 3, 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.).

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.

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?[23] 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 institutional obligations

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

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.[24]

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.

I am confronted with the following:

$$P \rightarrow (Q \rightarrow ((P \vee Q) \rightarrow \sim(\sim((P \vee P) \rightarrow P) \& \sim P \& \sim Q)))$$

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

[24] This fails to hold in cases where the statistical information is not within normal bounds, as in e.g. Williamson's (2000, p.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.

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.[25] 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 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

[25] 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.

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 by the simple argument.

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.[26] And given that such principles fail, it looks like appeals to

[26] 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

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.[27] 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, 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

---

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).

[27] Lasonen-Aarnio (2011). 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.

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.[28]

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.

[28] Given both her emphasis on unreasonable methods being bad for us to adopt (2011, section 5), and genuinely criticizable (2011, sections 5 & 7), (2014, p. 343), I suspect that Lasonen-Aarnio may ultimately have a picture like this. But then again, she also gestures toward an error theory; see ftn. 29.

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 a result, 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 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 thing 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.[29]

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.

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 a significant 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. Similarly, there will always be lingering worries that a purely counter-example driven case against a view may be vulnerable to the response that, yes, the view is ultimately

[29] 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. Aarnio (2011, p. 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.

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 8.

inconsistent with the highlighted intuitions about cases, but still, it is not the view that is thereby refuted. Rather, in the end it is the intuitions which we must give up.

To dispel this worry, my goal in this section is to sketch some central elements in a competing positive view—to speak to my own preferred understanding of the epistemic 'should.' I hold that this view can co-opt some of the motivations of right reason, while 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. In addition to casting light on *why* right reason is wrong, I also hope this alternative will go some way toward defusing a critique that my case rests *just* on (possibly suspect) intuitions.

As a preliminary matter, I should manage expectations: I am not aiming to conclusively defend—or even to fully state—my 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.

Begin with an overview. The thought animating the argument to follow will be that it is theoretically fruitful to analyze epistemic "should" judgments in terms of their relation to *epistemic self-regulation,* where this is conceived of as a distinctively epistemic cousin to ordinary practical planning. Just as we sometimes plan out *what action we should take* in a given situation, sometimes too do we plan out *what belief we should form* in a given situation. And in both cases our plans aren't just idle maps: the point of forming such a plan is to regulate our future behavior. The aim of regulation, though, comes with attendant constraints. Specifically, I claim that it is incoherent to form a plan that one believes can't, or won't, be able to guide one's behavior. But then, if we accept this, I think we have a good diagnosis of what's wrong with right reason. For when we try to translate the "should" judgments of right reason into plans, the natural candidates all look like they will be defective in just this way. Or, anyway, that's what I'll argue.

Start at the beginning—why believe that there is any connection between epistemic "should" judgments and epistemic self-regulation, as expressed in terms of *plans* or anything else? Well, here's

one pretty anodyne observation: in the ordinary course of things, our beliefs about what we *should* believe are at the very least 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. But, as indicated, I am going to pursue just one. Namely, I am going to proceed under the assumption 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 in some other way, 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. Given that this is so, what might 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.[30]

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 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*.[31]

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

[30] 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 (2015b). 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.'

[31] 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 8.

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 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.[32]

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.[33] Rather, in order to get right with rationality I would need to either change my beliefs about what I can discriminate (which 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.

---

[32] 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.

[33] 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.

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*.[34]

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 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 runnings 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.

[34] 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.

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 the 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.[35]

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.

---

[35] 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.

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 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.[36] Perhaps yet most modestly, one could think that expressability 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

[36] See Schoenfield (2015a) 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.

believe what the right reason view recommends.[37]

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.[38]

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*

[37] This is a good point to raise an issue on which I believe these stories may substantially diverge. I have argued above that the intelligibility of our epistemic planning depends, in part, on our beliefs about the efficacy of candidate epistemic plans. But these are just more beliefs—beliefs that themselves should be subject to epistemic norms. This is as of yet vague, but there are obvious regress worries lurking. My current inclination is to think that these worries are genuine and that they show a way in which epistemic self-regulation is inherently limited, and must, at some point, come to an end. In turn, I suspect that will be easier to accommodate on some of the theoretical options sketched in the main than others. But this issue is large, and although settling it would be very important to a fully developed account, it is nonetheless tangential to the diagnosis of right reason.

[38] An example of these resources in action: in Titelbaum's argument for an evaluatively-focused version of the right reason view (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. 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.

and indefeasible, and so on. We are now in a position to see 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.

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 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.[39] 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-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.[40] 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

[39] Hawthorne and Srinivasan (2013); for them, normative divergence is generally when the norms' recommendation comes apart from our judgments of blame, but blameworthy right-doing is the type relevant here.

[40] 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.

has really taught us. But adjudicating that issue is beyond the scope of this paper.[41] 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.[42]

Consider the sketch of the planning view offered in the last section. 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 planning view 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

[41] 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).

[42] 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.

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.

* * * *

Works Cited

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

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

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

---- (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.

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.

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 Szabó Gendler (eds.), Oxford Studies in Epistemology, Volume 1. Oxford University Press.

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

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

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

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

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

----, (2015a). Bridging Rationality and Accuracy. *The Journal of Philosophy*. 112 (12): 633–657.

----, (2015b) Internalism Without Luminosity. *Philosophical Issues.* 25 (1): 252–272.

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.

Titelbaum, Michael G. (2015). ‘Rationality’s Fixed Point (or: In Defense of Right Reason). In John Hawthorne & Tamar Szabó Gendler (eds.), Oxford Studies in Epistemology, Volume 5. Oxford University Press.

Weatherson, Brian. Do Judgments Screen Evidence?. Manuscript.

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

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

---- (2014). Very Improbable Knowing. *Erkenntnis*. 79 (5): 971–999.

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