Normative Externalism

Brian Weatherson

2019-05-19

Table of contents

Preface

Philosophy is hard. Ethics is hard; epistemology is hard; decision theory is hard; logic is hard. All the parts of philosophy are hard, but those four are going to be particularly relevant to the story I'm telling here. They matter because they are all evaluative. Someone who violates ethical principles is immoral; someone who violates epistemological principles is irrational; someone who violates the principles of decision theory is imprudent; someone who violates logical principles is illogical. And to say that someone is immoral, irrational, imprudent or illogical is to negatively evaluate them.

But it is easy to feel uneasy with this set of facts. If it is so hard to figure out the truth in these fields, why should we negatively evaluate someone for failing to conform to these hard to find standards? Doesn't fairness require that we only judge people by standards they can know about? I'm going to argue this is not right - that to evaluate someone is necessarily to impose a standard on them, and they may not even know what the standard is. Indeed, they may not have any reason to believe the truth about what the standard is, and in extreme cases may have good reason to endorse a false standard.

This position is uncomfortable, since it is easy to feel the unfairness of holding someone to a standard that they do not accept, and could not reasonably accept. Many philosophers think that we should either supplement or replace these external standards with internal standards. An 'internal standard' here is one that the person being evaluated either accepts, or has good reason to accept. To supplement the external standards is to say that there are two ways to evaluate people. It is good to live up to the correct standards in ethics, epistemology and decision theory, and bad to violate them. But it is also, say the supplementers, good to live up to one's own standards, and bad to violate them. The replacers say that conformity to one's own standards is more important than conformity to external standards; in some deep sense (at least some of) the heroes of ethics, epistemology and decision theory are people who abide by their own standards.

I am going to press two problems against this kind of view. The problems are

most pressing for the replacers, but they undermine the position of the supplementers too.

The first problem is that this kind of view has problems with fanatics and ideologues. Every ideologue who thought that they had figured out the one true way things must be done and reacted violently against those who didn't agree were doing well by their own lights. It's not good, in any way, to be that kind of ideologue. We shouldn't look back at the Reign of Terror and say, "Well, at least Robespierre and Saint-Just were living in accordance with their own values." Aiming to fit the world to one's own values is a dangerous game; it's only worth playing if you've got the values right. When we focus our attention on ideologues who have gone off the rails, the idea that it is unfair to hold people to a standard they can't see feels like something that's problem in theory but not in practice.

The second problem with the the internal view is that it leads to a nasty regress. It is, to be sure, hard to tell what the true values are. But choosing some values does not end our problems. Morality is hard even once you've settled on a moral theory. This is a point familiar from, for example, Sartre's discussion of the young man torn between duty to his mother and his country.

> What could help him make that choice? The Christian doctrine? No. The Christian doctrine tells us we must be charitable, love our neighbour, sacrifice ourselves for others, choose the "narrow way," et cetera. But what is the narrow way? Whom should we love like a brother–the solider or the mother? … Who can decide that *a priori*? No one. No code of ethics on record answers that question. (Sartre 1946/2007, 31)

We can evaluate the young man by his own lights and still be in a way unfair to him. Perhaps it turns out that the truly Christian thing to do is to fight Nazis, but the young man concludes (reasonably but falsely) that it is to help his mother. And he does that. If we are moved by the unfairness of holding him to a standard he does not endorse, we should also find it unfair to hold him to a consequence of his own standard that he doesn't recognise. But now what is left of the internal standard? It must be that it is good to do not what is best by one's own lights, but what one thinks is best by one's own lights. But perhaps one could even be wrong about *that*. (I'll discuss an example of this in chapter 1.) And the internal view collapses into the view that we should evaluate people by what they think they think they think … their own views support.

This is all absurd, and it makes the problem with fanatics and ideologues even worse. Perhaps we could argue that some ideologues take actions that are incompatible with what they say their values are. But they do not act against what they think their own values require.

Perhaps we can motivate the importance of the internal point of view not by thinking about fairness, but by focussing on an analogy with reckless agents. If I fire a cannon down Fifth Avenue at peak hour, I do something morally horrible even if miraculously I don't hit anyone. My action is wrong because it is reckless. Perhaps if I do something that is probably morally wrong, I am morally reckless in just the same way. And that's true even if my action turns out not to be wrong. So what matters is not just what is right and wrong, but probabilities of rightness and wrongness. I think this kind of reasoning fails too, and there are important asymmetries between physical risk (as is involved in firing cannons down busy streets) and moral risk. I'll spend chapters three and four outlining these asymmetries, and why they tell against the idea that there is a distinctive wrong of moral recklessness.

The first half of the book discusses the significance of the internal point of view in ethics. As I've indicated, I don't think it is particularly important, though we'll spend a bit of time towards the end of part one looking at some more limited, and hence more plausible, claims for its usefulness. The second part of the book turns to epistemology, and to the idea that one cannot reasonably have beliefs that one believes (or should believe) to be unreasonable.

Again, the issue turns on how important is conformity to one's own standards. The most common philosophical view around here is a kind of supplementing view, not a replacing view. It is important, say several philosophers, to have beliefs that are both actually reasonable and also reasonable by one's own lights. And I'm going to push back against that. One reason comes from work by Timothy Williamson. What's reasonable to believe turns on empirical facts about one's situation. Since we don't have God-like perfect access to our own empirical situation, we might not realise what is reasonable to do in our own situation just because we don't know precisely what situation we are in. In such cases, it seems we should react to the situation we are actually in, not to our best guess about what situation that is.

There will be two primary themes of part two of the book. One echoes the first part of the book. Sometimes we cannot know what it would be to be reasonable by our own lights. So adding a requirement that reasonable people are doing well by their own lights threatens to trigger a vicious regress. I'm going to argue that

this threat is realised. The other theme is that the phenomena that philosophers have thought could only be explained by adding an internal constraint onto belief can be adequately explained by a more careful attention to the nature of evidence, and what it takes for one to have evidence and for that evidence to support a belief. I'll argue that such explanations are preferable to explanations in terms of internal constraints (such as only believe what you believe is reasonable to believe). This is in part because they avoid regress and implausible knowledge about one's own situation; in part because they only commit us to things we are independently committed to; and in part because they explain a much broader range of cases than are explained by the alleged internal constraints.

I have more people to thank for help with this book than I could possibly list here. I'm not even sure at which point of time I should start the thanks. Twenty-odd years ago as a graduate student at Monash I wasn't working on *this* project. But the picture that pervades this book, that in philosophy everything is contestable and there are no safe stopping points, owes a lot to the amount of time I spent as a graduate student thinking about, and being taught about, heterodox approaches to logic and to decision theory.

Most of the best feedback I've received on the various parts of the book has come from graduate students. Some of the second part of the book is based on an epistemology seminar I taught at Rutgers. I taught a graduate seminar at Michigan off an early draft of the book manuscript. And I've taught several mini-courses at St Andrews, and presented at even more workshops and symposia there, off parts of the book. In every case the feedback I received from colleagues and, even more frequently, graduate students, changed the book for the better.

Parts of the book are based on presentations at or organised by the University of Aberdeen, University of Oxford, University of Vienna, University of Konstanz, University of Zurich, University of Graz, Massachusetts Institute of Technology, Princeton University, Ohio State University, University of Sydney, Australian National University and University of Melbourne. I've presented parts of it at the Bellingham Summer Philosophy Conference, the Night of Philosophy in New York City and the Australasian Association of Philosophy annual conference. And I've discussed it with the Corridor Reading Group in New York, and the Ethics Lunch group in Ann Arbor. I'm very grateful for all the feedback I got at those presentations.

As well as all those audiences, I'd like to particularly thank Derek Ball, Jessica Brown, Sarah Buss, Herman Cappelen, Ruth Chang, Stewart Cohen, Josh Dever, Tom Donaldson, Andy Egan, Claire Field, Katherine Hawley, Scott Hershowitz,

Torfinn Huvenes, Jonathan Jenkins Ichikawa, Jim Joyce, Zoe Johnson King, Maria Lasonen-Aarnio, Ben Levinstein, Julia Markovits, Matthew McGrath, Sarah Moss, Jill North, Caroline Perry, Quentin Pharr, Lewis Ross, Andrew Sepielli, Joe Shin, Holly Smith, Martin Smith and Elia Zardini for particularly valuable feedback. (And I'm already dreading finding out who I should have included on this list but didn't.) Ralph Wedgwood read the whole manuscript and provided comments that improved it in innumerable ways. Thanks to him, and to Peter Momtchiloff for making such an astute choice of reader for the manuscript.

The idiosyncratic workflow I used for writing this would have been impossible without Fletcher Penney's Multimarkdown (both the language and the Composer software) and John MacFarlane's Pandoc, and I'm very grateful to both of them for building such valuable tools. Much of the book was drafted under the dome in the La Trobe Reading Room at the State Library of Victoria, and I'm so grateful that Victoria has maintained that space, and that building.

Early in the development of this book project, I was honoured to become the first Marshall M. Weinberg Professor of Philosophy at the University of Michigan, Ann Arbor. Without the support Marshall has provided to my research, and to the research project of the University of Michigan more broadly, this project would have been unimaginable. My inaugural lecture was "Running Risks Morally", most of which appears in one way or another in part one of the book. The first draft of the book was written while on a sabbatical funded through the Weinberg Professorship. But beyond that, the vibrant intellectual community here at Michigan relies in ever so many ways on Marshall's support. I couldn't tell you how much this book relies on feedback from graduate students who have received Weinberg fellowships, or who came to Michigan in part because of the Weinberg Center for Cognitive Science. While this is by no means a work of cognitive science, it is influenced in many ways by what I've learned from cognitive scientists talking at the Weinberg Center. And I really cannot thank Marshall enough for his support for Michigan, and for its research.

Finally, I'd like to thank Ishani Maitra and Nyaya Maitra Weatherson for, well, everything. Ishani didn't just talk through all the things in this book with me, and improved it in so many ways, but she also talked through all the things I cut from the book. And she improved those portions too.

1. Introduction

1.1. To Thine Own Self Be True

Early in *Hamlet*, Laertes departs Elsinore for Paris. As he prepares to go his father, Lord Polonius, offers him some paternal advice. He tells him to talk less and smile more. He tells him to spend all his money on clothes, since that's how they roll in Paris. He tells him to neither a borrower nor a lender be, though the latter is presumably redundant if he's taken the advice to date. And he concludes with this advice, destined to adorn high school yearbooks for centuries to come.

> This above all: to thine own self be true,
> And it must follow, as the night the day,
> Thou canst not then be false to any man.

It isn't completely clear what Polonius means when he advises Laertes to be true to himself, but it is plausible that he means something like this:

> Follow your own principles!

Or perhaps something like this:

> Do what you think is right!

And unlike the rest of the advice Polonius gives, many philosophers have followed him in thinking this is a very good idea.

The primary aim of this book is argue against this idea. Following one's own principles, or doing what one thinks is right, are not in general very good ideas at all. I will call normative internalism the view that we should be guided by norms that are internal to our own minds, in the sense that our beliefs, and our (normative evidence) is internal to our minds. And I will oppose that view, arguing for normative externalism.

Normative externalism is the view that the most important standards for evaluating actions, mental states and agents are typically external to the actor, believer or agent being evaluated. It can be appropriate to hold someone to a moral, or an epistemic, standard that they do not endorse, or even that they could not be reasonably expected to endorse. If one has bad standards, there need be nothing wrong in violating them, and there is nothing good about upholding them.

That last paragraph made a lot of distinct claims, and it is worth spending some time teasing them apart. But before we get too deep in the weeds, I want to have on the table the guiding principle of the book. Being true to yourself, in the sense of conforming to the principles one has, or even to the principles one has reason to have, is just not that important. What is important is doing the right thing, being a good person, and having rational beliefs. If one has misguided views about the right, the good, and the rational, then there is nothing good about conforming to those misguided views. And this matters, because many people have views about the right, the good, and the rational, that are very misguided indeed.

1.2. Four Questions

1.2.1. Actions, Agents or Advice

If one says, with Polonius, that it is good to conform to one's own principles, there are a number of distinct things one could be meaning.

One could be making a claim about particular *actions*. (Or about particular beliefs, but we'll focus on actions for the next few paragraphs.) So one could be saying that actions that conform to the actor's principles are thereby in some sense right or good, and those that violate the actor's principles are in some sense wrong or bad.

Alternatively, one could be making a claim about *agents*. So one could be saying that people who (typically) conform their actions to their principles are in some sense good (or less bad) people, and those who violate their own principles are in some sense bad.

Or alternatively again, one could be making a claim about *advice*. One could be saying that whether or not the claims in the previous two paragraphs are strictly correct, it is excellent to advise people to act according to their principles. There

are plenty of cases where advising people to do the optimal thing is bad, especially if aiming for the optimal result is likely to lead to catastrophe. So this view about advise is in principle distinct from the views about actions and agents.

The form of externalism I will defend is opposed to the views in all of the last three paragraphs. But it is most strongly opposed to the view about actions, and least strongly opposed to the view about advice. Indeed, I won't have a lot to say about advice throughout the book; except to note occasionally when intuitions about advice seem to be getting used illegitimately to justify conclusions about actions. But I don't mean to imply that the views have to stand or fall together. A view that is externalist about actions - it thinks it doesn't make any difference to the correct evaluation of an action whether the actor endorsed it or not - but internalist about agents - it thinks there is something good about people who stick to their principles and bad about those who do not - is certainly worth considering. But it isn't my view; I mean to oppose all three precisifications of what Polonius says.

1.2.2. Above All?

Polonius does not just say Laertes should be true to himself. He says this is something 'above all'. This suggests that he is elevating *Do what you think is right* to a central place, making it more important than principles like *Respect other people*, or *Make the world better*, or even *Do the right thing*.

The externalist view I support takes completely the opposite tack. The principle *Do what you think is right* is of no importance at all.

But there is a large middle ground position. This is easiest to see if we assume the debate is about agents, not actions or advice, so I'll present it for agents. But it shouldn't be too hard to see how to generalise the idea.

We could hold that doing what one thinks is right is one of the virtues, something that contributes to a person being a good person. Or we might think that failing to do what one thinks is right is a vice, something that contributes to a person being a bad person. And we might think one or other (or both) of those things without thinking them particularly important virtues or vices. One could coherently hold that there is a virtue in holding to one's principles, even if one thinks that other virtues to do with honesty, courage, respect and the like are more important. And one could coherently hold that doing what one thinks is

wrong is a vice, even in the case where one has false enough views about first-order moral questions that doing what one thinks it right would manifest even more serious vices.

Indeed, one might think that ordinary English goes along with this. We do talk somewhat admiringly about people who are principled or resolute, and somewhat disdainfully about people who are hypocritical.[1]

I'm going to classify this kind of view, the one that says that doing what one thinks is right is important to character, but not of maximal importance, as a moderate internalist view. And my externalism will be opposed to it, like it is opposed to the view that being principled, and avoiding hypocrisy, are the most important virtues.

The possibility of such a moderate internalist view is important, because otherwise we might think the argument against internalism would be too easy. History is full of fanatics who convinced themselves that they were doing the right thing while causing immense harm. It is hard to believe that the one principle they did conform to, *Follow your own principles*, is the most important principle of all. But perhaps, just perhaps, their resoluteness is in a small way a virtue. At least, a philosophical view that says that it is a virtue, albeit one offset by mountains of vice, is not absurd.

1.2.3. Ethics, Epistemology and More

I've been interpreting Polonius's dictum as being primarily about ethics so far. But views like his are available in many other areas of philosophy. I'll mention three more here, the first of which will be a major focus of this book.

Belief is subject to evaluation on a number of fronts. Beliefs are true or false, but that hardly exhausts their virtues or vices. Some true beliefs are bad in virtue of being lucky guesses, or leaps to unwarranted conclusions. Some false beliefs are the result of sensibly following the evidence where it leads, and just being unluckily misled into error. So as well as evaluating a belief for truth, we can evaluate it for responsiveness to the evidence. I'm going to argue, somewhat

[1] Though to be clear, I don't think the English words 'principled' and 'resolute' actually pick out the so-called virtue of upholding one's own principles. Following Richard Holton (1999), I think those words pick out diachronic properties of a person. They apply to a person in part due to that person's constancy over time in some respect. Following one's principles isn't like this; it is a purely synchronic affair.

indirectly, that a belief is rational just in case it is responsive to the evidence in this way.[2]

But if that's what rationality is, then subjects can also have beliefs about the rationality of their own beliefs. And we can ask whether subjects are doing well at believing by their own lights. To believe something just is to believe it is true, so if our only standard for belief is truth, then everyone will believe well by their own lights. But it is possible to believe something, and even rationally believe it, while believing that that very belief is irrational. Or, at least, so I'll argue.

Is this a bad thing? Should we mark someone down for believing in a way that they take to be irrational? I'm going to argue that we should not. It's good to believe truths. It's good to believe in accord with one's evidence. And that's as far as we should go. It's not good to believe in accord with what one believes the evidence supports, unless one thereby ends up with a belief that is good for some other reason. And it's not bad to believe something that one believes is not supported by one's evidence, unless one ends up with a belief that is bad for some other reason.

Just as in the ethics case, we can separate out a number of distinct questions here. Assume you think there is something philosophically important about beliefs that are irrational by the lights of the believer themselves. You could say that this is a bad-making feature of the belief itself, or a bad-making feature of the believer, or, perhaps that it is bad to advise people to have beliefs that are irrational by their own lights. That is, we can replicate the act, agent or advice distinction inside epistemology, though the 'acts' are really the states of holding particular beliefs. And if you do think these beliefs, or believers, are bad in some way, there is a further question about how much badness is involved. Is believing in a way that one thinks is irrational as bad as not following the (first-order) evidence, or more bad, or less bad. (Or is badness the wrong concept to be using here?)

We will see different philosophical views that take different stands on these questions throughout part II of the book. I'm going to defend a fairly simple, and fairly extreme, position. It isn't a bad making feature, in any way, of a belief that the believer thinks it is irrational, nor is it a bad making feature of believers that they have beliefs they think are irrational. It isn't even a bad habit to routinely

[2]Though getting clear on just what this last sentence commits me to will require saying more about what evidence is. For now, it won't do much harm to equate evidence with basic knowledge. A proposition p is part of the subject's evidence if the subject knows p, and doesn't know p because she inferred it from something else.

have beliefs that one thinks are irrational; though I'm going to be a little more tentative in defending that last conclusion. The general principle throughout is to motivate and defend a picture where what matters is conformity to the actual rules - be they rules of action or rules of belief - rather than conformity to what one takes (or even rationally takes) the rules to be.

The disputes of the last few paragraphs have all been over epistemology, fairly narrowly construed. But there are some other disputes that we might have to, where the difference between conformity to external rules and conformity to one's version of the rules matters. I'm not going to say much about the next two disputes, but they are helpful to have on the table.

Some lives go better than others. When we act for the sake of others, when we act benevolently, we aim to improve the lives of others. Call someone's *welfare* that quantity we improve when we act benevolently.[3] Philosophers disagree a lot about what welfare is, so some of them are wrong. And though I'm not going to argue for this, it seems to me that the disagreeing parties each have such good arguments that at least some of the philosophers who are wrong are nevertheless rational in holding the position they do. So that implies that a rational person could have a choice between two actions, one of which actually produces more welfare, and the other of which produces more welfare according to the theory of welfare they (rationally) hold. Assuming the person wants to act benevolently, or, if the act is directed to their own good, they want to act prudentially, is there anything good about doing the thing that produces more welfare according to the theory of welfare they hold? My position, though I'm not going to argue for this in this book, is that there is not. What matters for benevolent or prudential action is how well one's act does according to the correct theory of welfare. It doesn't make an action benevolent, or prudent, if the action is good according to a mistaken theory of welfare. That's true even if the theory of welfare is one's own, or even if it is the one that is rational for one to hold. If one's theory of welfare is a purely hedonistic experiential theory of welfare, then you might think you are improving the welfare of others by force-feeding them happy pills. But if that theory of welfare is false, and welfare involves preference satisfaction, or autonomy, then such an action will not be benevolent, nor will it be rational to perform on benevolent grounds.

We can make the same kind of distinction within decision theory. Let's assume for now that a person has rational beliefs, and when they lack belief they assign a rational probability to each uncertain outcome, and they value the right

[3]There are a lot of different things that people call welfare in the philosophical literature. I'm taking the idea of tying it definitionally to benevolent action from Simon Keller (2009).

things. There is still a question about how they should act in the face of uncertainty. Unlike the questions about ethics, epistemology, or welfare, there is an orthodox answer here. They should maximise expected utility. That is, for each act, they should multiply the probability of each outcome given that act, by the (presumably numerical) value of that outcome-act pair, and add up the resulting products to get an expected value of the act. Then they should choose the act with the highest expected value. But while this is the orthodox view of decision theory, there are dissenters from it[4]. The best recent statement of dissent is in a book-length treatment by Lara Buchak (2013). And someone who has read Buchak's book can think that her view is true, or, perhaps, think that there is some probability that it is true and some probability that the orthodoxy is true.

So now we can ask the same kind of question about conformity to the correct rules versus conformity to the rules one thinks are correct.[5] Assume that someone does not have the correct beliefs about how to rationally make decisions. And assume that they perform an act which is not rational, according to the true decision theory, but is rational according to the decision theory they accept. Is there something good about that decision, and would there have been something bad about them doing the thing that correct theory recommended? My position is that there is not. The rational decisions are the ones recommended by correct decision theory. There is nothing to be said for conforming to one's own preferred decision theory, if that theory is false.

1.2.4. Actual or Rational

So far I've focussed on the distinction between principles that are external to the agent, and principles that are internal to the agent in the sense of being believed by the agent, or being the agent's own principles. When I call my view externalist, it is to indicate that I think it is the external principles that matter. But there is another category of principles that I haven't focussed on, and which are in some sense internal. These are the principles that the agent should, rationally, accept.

[4] I'm suppressing disputes within orthodoxy about how just to formulate the view, though those disputes would also suffice to get the kind of example I want going.

[5] If the moral theories one gives credence to reject expected value maximisation, then there will be even more complications at the intersection of ethics and decision theory. Ittay Nissan-Rozen (2015) has a really nice case showing the complications that arise for the internalist when moral theories do not assume orthodox decision theory.

Now if we say that the agent should rationally accept all and only the true principles, then there won't be a distinction between *Follow the true principles* and *Follow the principles it is rational to accept*. But let's work for now with the assumption that there is a difference here; that just like with anything else, agents can be rationally misled about the nature of ethics, epistemology, welfare, and decision theory.[6] Then there is another possibility; that agents should follow the principles that they have most reason to believe are true.

This gives another way for the internalist to respond to the problem of historical monsters. Let's think about one particular case, one that I'll return to occasionally in the book: Maximilien Robespierre[7]. Whatever else one can say about him, no one seriously doubts that Robespierre always did what he thought was right.[8] But doing what he thought was right involved setting off the Reign of Terror, and executing ever so many people on incredibly flimsy pretexts. We can't really say that the principle he did well by, *Do what you think is right*, is one that should be valued above all. We mentioned above that we could reasonably say it is a good-making feature of Robespierre that he was principled, even if it is outweighed by how abhorrent his set of principles turned out to be. But the interest here is in whether we can find some internalist principle that can be said to be true 'above all' in his case.[9]

Robespierre had ample reason to believe that he had ended up on the wrong

[6] Julia Markovits (2014) argues that agents have rational reason to accept the fundamental moral truths. Michael Titelbaum (2015) argues that agents have rational reason to accept the fundamental epistemological truths. I'm assuming for now that both of these positions are false, because it gives my opponents more room to move if they are false. Claire Field (forthcoming) responds to Titelbaum's arguments. Note here that when I say that an agent can be rationally misled about morality and epistemology, I am not claiming that they can rationally have outright false beliefs about morality and epistemology. I just mean that rationality is consistent with having something other than complete certainty in the claims that are actually true.

[7] There are more historical sources on Robespierre than would be remotely possible to list. The things I say here are largely drawn from recent work by Peter McPhee (2012), Ruth Scurr (2006) and especially Marisa Linton (2013). The study of the Committee of Public Safety by R. R. Palmer (1941) is helpful for seeing Robespierre in context, and especially seeing him alongside men with even more extreme characteristics than his.

[8] Most revolutionary leaders are either power-hungry or bloodthirsty. But Robespierre genuinely seems to have been neither of those, except perhaps at the very very end. Linton (2013, 97–99) is particularly clear on this point.

[9] One thing that won't rescue intuitions about the case is to say that *Do what you think is right* is important only if the agent is 'procedurally rational'. Robespierre used the right methods to form moral beliefs: he read widely, talked to lots of people, and reflected on what he heard and saw. He just got things catastrophically wrong. Gideon Rosen (2003, 2004) places a lot of emphasis on procedural rationality in defending a form of internalism, though his aim is very much not to track intuitions about particular cases.

track. He wasn't brainwashed into believing that the Terror was morally justifiable; the reasons for it were clearly present to him. The results of the Terror weren't playing out in some distant land, or in the hold of a slave ship, they were right in front of him. And he knew a lot of moral and political theory. He was well educated in the classics. He read Montesquieu. He read, and adored, Rousseau. He sat through hours upon hours of debate every day about the efficacy and morality of government actions, both before and during his reign. Even if one thinks, as I do, that sometimes the reasons for the immorality of an action are hidden from the actor, that can hardly be said to be true in Robespierre's case.

So I think we can reasonably say in Robespierre's case that he violated the rule *Follow the principles it is rational to accept*. And that rule is an internal rule, in some sense. If we take it to be the primary rule, then we won't judge people by standards that are hidden from them. We may judge them by standards they don't accept, but only when they have reason to accept the standards. So I'll treat it as another internalist approach, though very different from the approach that says it is most important for people to follow their own principles.

So we have two very different kinds of internalist approaches to ethics, epistemology, welfare and decision theory. One says that it is (most) important that people follow their own principles. The other says that it is (most) important that people follow the principles they have rational reason to accept. The first, in its strongest form, says absurd things about the case of fanatics. As I'll argue at length in what follows, it also leads to nasty regresses. The second does not have these problems. But it is very hard to motivate. We will spend some time on the reasons philosophers have had for wanting views like Polonius's. All of these, I'll argue, push towards the idea that the most important thing is that people follow the principles they actually accept. None of them, when considered carefully, give us a reason to prefer principles the actor or believer has reason to accept to the principles that are actually true. Retreating from *Follow your own principles* to *Follow the principles it is rational to accept* lets the internalist avoid harsh cases like Robespierre, but at the cost of abandoning the interesting reasons they have for their view.

1.2.5. Some Caveats

I've spoken freely in this section about the true moral principles. That way of speaking presupposes that there are moral truths. I mean to be using the phrase 'moral truths' in as non-committing as sense as is possible. I don't mean to say

that the moral truths are mind-independent. If it is true that murder is wrong in virtue of our disapproval of murder, it is still true that murder is wrong, and that's enough for current purposes. Nor do I mean to insist that the moral truths are invariant across space and time. There are hard questions about how we should evaluate actors from different times and places if a form of moral relativism is true. But those questions are largely orthogonal to the one's I'm interested in.

I am in effect assuming away a very strong form of moral relativism, one that makes moral truth relative to the moral principles of the actor being evaluated. But that's not a plausible form of moral relativism. If moral relativism is true, then what morality is relative to is much more inclusive than a single person; it is something like a culture, or a practice. And that is enough for there to be a difference between what a person accepts, and what is true in their culture or practice.

As briefly noted above, I'm also assuming that there is a difference between what is true and what it is rational to accept. All I really need here is that it can be rational to be less than fully certain in some moral and epistemic truths. I'm not going to assume, for example, that one can rationally believe moral or epistemic falsehoods. I've spoken above as if that is possible, but that was a convenient simplification. What's going to really matter is just the existence of a gap between what's true and what's reasonable to believe, and that gap can arise even if all the things that are reasonable to believe are true.

Finally, you may have noticed that we ended up a long way from anything that could be plausibly attributed to Lord Polonius. When he tells Laertes to be true to himself, I'm pretty sure he's not saying anything about whether Laertes should have beliefs that are rational by the standards that Laertes should rationally accept. Yet whether Laertes (or anyone else) should have such beliefs is one of the questions we ended up being interested in. The good Lord's role in this play was just to introduce the distinction between following one's own principles and following the true principles. With that distinction on stage, we can let Polonius exit the scene.

1.3. Normative Externalism Defined

Normative externalism is the view that the most important evaluations of actions and actors, and of beliefs and believers, are independent both of the actor or believer's belief about the value of their action or belief, and of the evidence the

actor or believer has about the value of their action or belief. The aim of this book is to defend normative externalism, and explore why it is philosophically important.

It is tempting to strengthen this kind of normative externalism further, and say that what one should do and believe is completely independent of what one believes one should do and believe. But this strong independence claim can't be right. (I'm grateful here to Derek Ball.) If one thinks that one should murder one's neighbours, then one ought to get professional help. Sometimes normative beliefs change the normative significance of other actions. So the externalist claim I'm defending is a little weaker than this general independence claim. It allows that a normative belief B may change the normative status of actions and beliefs that are not part of the content of B. But the externalism I'm defending is still going to be strong enough to allow a lot of critics.

The strongest kind of normative internalism says that the value of actions and beliefs is tightly tied to the beliefs that actors and believers have about their own actions and beliefs. It says that the most important moral precept is to do what you think is right, and the most important epistemological precept is to believe what you think the evidence supports. The strong version of internalism is not a popular position. But it has an important role to play in the narrative here. That's because there are many interesting, and popular, moderate versions of internalism. Yet once we look at the motivations for those moderate versions, we'll see that they really are arguments for the strongest, and least plausible, version.

We can generate those moderate forms of normative internalism by looking at the four questions from the previous section. Some internalists say that internalism is true just for actors (or believers), not for actions (or beliefs). Some say that internalist principles are part of the moral (or epistemological) truth, not principles to put above all. Some say that internalism principles apply to just one of ethics or epistemology, not both. And some say that what matters is not conformity to the principles one actually holds, but conformity to the principles one has evidence for. And answers to these questions can be mixed and matched indefinitely to produce varieties of internalist theses. Here, for example, are three principles that are both widely believed, and which you can get by mixing and matching answers to the four questions.

- It is a vice to frequently do things one believes are wrong, even if those actions are actually right.

- Wrong actions are blameless, and hence do not reflect badly on the actor who performs them, if that actor believes the action is right, and has good reason for that belief.
- A belief is irrational if the believer has good evidence that the belief is not supported by their evidence, even if that 'higher-order' evidence is misleading.

And I'm going to argue that the best arguments for those positions overgeneralise; they are equally good as arguments for the implausible strong version of internalism. So they are no good.

Part of the argument here will be piecemeal: showing for a particular internalist thesis that there are no good arguments for it but for the arguments that lead all the way to the strongest form of internalism. And I can't hope to do that for all the possible theses you could get by mixing and matching answers to the four questions. But I can hope to make the strong form of externalism more plausible, both by showing how it handles some difficult cases, and by showing that the most general arguments against it do not work.

1.4. Guidance

To illustrate the kind of storyline I sketched in the previous section, let's consider one popular argument against externalism. The externalist says that people should do the right thing, whatever that is, whether or not they know that the right thing is in fact right. It is often objected that this is not particularly helpful guidance, and morality should be more guiding than this. We see versions of this objection made by Ted Lockhart (2000, 8–9), Michael M. Smith (2006, 143), Andrew Sepielli (2009, 8), William MacAskill (2014, 7) and by Hillary Greaves and Toby Ord (2017). These authors differ between themselves about both why norms that are not guiding are bad, some saying they are unfair, others that they are unhelpful, and about what conclusion we should draw from this fact. But they agree there is something bad about *Do the right thing* in virtue of it not being guiding, and think we need something more internalist.

But if you think *Do the right thing* is not guiding, and we need norms that are guiding in just that sense, some very strong conclusions follow. After all, if non-guiding rules are bad, then they shouldn't be any part of our moral theory. So it isn't just that we should take hypocrisy to be one vice alongside cowardice, dishonesty, and so on, but to be the only vice. After all, if there are other vices

at all, then morality as a whole may not be guiding. Now who is *Do the right thing* not guiding to? Presumably to people who lack full moral knowledge. But some of these people won't have full epistemological knowledge either. So by the standard that *Do the right thing* is not guiding, principles like *Do whatever the evidence best suggests is right*, or *Do whatever maximises expected rightness* won't be guiding either. If we can't expect people to know what's right, we can't really expect them to know what's probably right either.

So taking guidance to be a constraint in this way pushes us to a version of internalism that relies on actual beliefs about rightness, not beliefs the evidence supports, and relies on a version that takes conformity to one's own values to be 'above all'. But if we do that, we can't say either of the plausible things I suggested various moderate internalists could say about Robespierre. The two suggestions were to say that conformity to one's own value is merely one virtue among many, and that good people should conform not to their actual principles, but to the principles their evidence supports. If we take guidance to be a constraint, then both ways out are blocked. Robespierre failed by some very important standards, but he couldn't be guided (in whatever sense the internalist means) by those standards.

We'll see this storyline a few times in what follows. The externalist view seems to have some unattractive features. But when we spell out just what the features are, we'll see they are shared by all but some very implausible theories. This won't just hold in ethics. The epistemological picture I'm going to draw allows for kinds of reasoning that appear on their face to be unacceptably circular. But when we try to say just what this kind of circularity comes to, we'll see that blocking it would provide enough resources to ground an argument for Pyrrhonian scepticism.

1.5. Symmetry

In general, one's evidence is relevant to what one should do. The normative externalist denies a natural generalisation of this little platitude. Although evidence about matters of fact is relevant to what one should do, evidence about the normative, about the nature of morality and rational, is not. Evidence about whether to turn left or right is relevant to rational decision making, evidence about what is wrong or right is irrelevant. Or so says the externalist.

This looks like an argument against externalism: it denies a very plausible symmetry principle. The principle says that we should treat all kinds of uncertainty, and all kinds of evidence, the same. I'm going to spend much of the first half of this book arguing against the symmetry principle, but for now let's quickly set up why we might think there is a puzzle here.

We'll start by thinking through an example of where evidence is relevant to mundane action. A person, we'll call him Baba, is looking for his car keys. He can remember leaving them in the drawer this morning, and has no reason to think they will have moved. So the natural thing to do is to look in the drawer. If he does this, however, he will be sadly disappointed, for his two year old daughter has moved the car keys into the cookie jar.

Things would go best for Baba if he looked in the cookie jar; that way he would find his car keys. But that would be a very odd thing for him to do. It would be irrational to look there. It wouldn't make any sense. If he walked down the steps, walked straight to the cookie jar, and looked in it for his car keys, it would shock any onlookers because it would make no sense. It used to be thought that it would not shock his two year old daughter, since children that young had no sense that different people have different views on the world. But this isn't true; well before age two children know that evidence predicts action, and are surprised by actions that don't make sense given a person's evidence (He, Bolz, and Baillargeon 2011). This is because from a very young age, humans expect other humans to act rationally (Scott and Baillargeon 2013).

In this example, Baba has a well-founded but false belief about a matter of fact: where the car keys are. Let's compare this to a case where the false beliefs concern normative matters. The example is going to be more than a little violent, though after this the examples will usually be more mundane. And the example will, in my opinion, involve three different normative mistakes.

> Gwenneg is at a conference, and is introduced to a new person. "Hi," he says, "I'm Gwenneg," and extends his hand to shake the stranger's hand. The stranger replies, "Nice to meet you, but you shouldn't shake my hand. I have disease D, and you can't be too careful about infections." At this point Gwenneg pulls out his gun and shoots the stranger dead.

Now let's stipulate that Gwenneg has the following beliefs, the first of which is about a matter of fact, and the next three are about normative matters.

First, Gwenneg knows that disease D is so contagious, and so bad for humans both in terms of what it does to its victims' quality and quantity of life, that the sudden death of a person with the disease will, on average, increase the number of quality-adjusted-life-years (QALYs) of the community.[10] That is, although the sudden death of the person with the disease obviously decreases their QALYs remaining, to zero in fact, the death reduces everyone else's risk of catching the disease so much that it increases the remaining QALYs in the community by a more than offsetting amount.

Second, Gwenneg believes in a strong version of the 'straight rule'. The straight rule says that given the knowledge that $x\%$ of the Fs are Gs, other things equal it is reasonable to have credence that this particular F is a G. Just about everyone believes in some version of the straight rule, and just about everyone thinks that it needs to be qualified in certain circumstances. When I say that Gwenneg believes in a strong version of it, I mean he thinks the circumstances that trigger qualifications to the rule rarely obtain. So he thinks that it takes quite a bit of additional information to block the the transition from believing $x\%$ of the Fs are Gs to having credence that this particular F is a G.[11]

Third, Gwenneg thinks that QALYs are a good measure of welfare. So the most beneficent action, the one that is best for well-being, is the one that maximises QALYs. This is hardly an uncontroversial view, but it does have some prominent defenders (McKie et al. 1998).

And fourth, Gwenneg endorses a welfarist version of Frank Jackson's decision-theoretic consequentialism (Jackson 1991). That is, Gwenneg thinks the right thing to do is the thing that maximises expected welfare.

Putting these four beliefs together, we can see why Gwenneg shot the stranger. He believed that, on average, the sudden death of someone suffering from disease D increases the QALYs remaining in the community. By the straight rule, he inferred that each particular death of someone suffering from disease D increases the expected QALYs remaining in the community. By the equation of QALYs with welfare he inferred that each particular death of someone suffering from disease D increases the expected welfare of the community. And by his welfarist consequentialism, he inferred that bringing about such a death is a good

[10] QALYs are described in McKie et al. (1998), who go on to defend some philosophical theses concerning them that I'm about to assign to Gwenneg.

[11] Nick Bostrom (2003) endorses, and uses to interesting effect, what I'm calling a strong version of the straight rule. In my reply to his paper I argue that only a weak version is plausible, since other things are rarely equal (Weatherson 2003). Gwenneg thinks that Bostrom has the better of that debate.

thing to do. So not only do these beliefs make his action make sense, they make it the case that doing anything else would be a moral failing.

Now I think the second, third and fourth beliefs I've attributed to Gwenneg are false. The first is a stipulated fact about the world of Gwenneg's story. It is a fairly extreme claim, but far from fantastic. There are probably diseases in reality that are like disease D in this respect[12]. So we'll assume he hasn't made a mistake there, but from then on every single step is wrong. But none of these steps are utterly absurd. It is not too hard to find both ordinary reasonable folk who endorse each individual step, and careful argumentation in professional journals in support of those steps. Indeed, I have cited just such argumentation. Let's assume that Gwenneg is familiar with those arguments, so he has reason to hold each of his beliefs. In fact, and here you might worry that the story I'm telling loses some coherence, let's assume that Gwenneg's exposure to philosophical evidence has been so tilted that he has only seen the arguments for the views he holds, and not any good arguments against them. So not only does he have these views, but in each case he is holding the view that is best supported by the (philosophical) evidence available.

Now I don't mean to use Gwenneg's case to argue against internalism. It wouldn't be much use in such an argument for two reasons. First, there are plenty of ways for internalists to push back against my description of the case. For example, perhaps it is plausible for Gwenneg to have any one of the the normative beliefs I've attributed to him, but not to have all of them at once. Second, not all of the internalist views I described so far would even endorse his actions given that my description of the case is right.

But the case does illustrate three points that will be important going forward. One is that it isn't obvious that the symmetry claim above, that all uncertainty should be treated alike, is true. Maybe that claim is true, but it needs to be argued for. Second, the symmetry claim has very sweeping implications, once we realise that people can be uncertain about so many philosophical matters. Third, externalist views look more plausible the more vivid the case becomes. It is one thing to say abstractly that Gwenneg should treat his uncertainty about morality and epistemology the same way he treats his uncertainty about how many people the stranger will infect. At that level of abstraction, that sounds plausible. It is another to say that the killing was a good thing. And we'll see this pattern a lot as we go forward; the more vivid cases are, the more plausible the externalist

[12] At least, there probably were such diseases at some time. I suspect cholera had this feature during some epidemics.

position looks. But from now on I'll keep the cases vivid enough without being this violent.[13]

1.6. Regress

In this book I'm going to focus largely on ethics and epistemology. Gwenneg's case illustrates a third possible front in the battle between normative internalists and externalists: welfare theory. There is a fourth front that also won't get much discussion, but is I think fairly interesting: decision theory. I'm going to spend a bit of time on it right now, as a way of introducing regress arguments for externalism. And regress arguments are going to be very important indeed in the rest of the book.

Imagine that Llinos is making trying to decide how much to value a bet with the following payoffs: it returns £10 with probability 0.6, £13 with probability 0.3, and £15 with probability 0.1. Assume that for the sums involved, each pound is worth as much to Llinos as the next.[14] Now the normal way to think about how much this bet is worth to Llinos is to multiply each of the possible outcomes by the probability of that outcome, and sum the results. So this bet is worth $10 \times 0.6 + 13 \times 0.3 + 15 \times 0.1 = 6 + 3.9 + 1.5 = 11.4$. This is what is called the *expected* return of the bet, and the usual theory is that the expected return of the bet is its value. (It's not the most helpful name, since the expected return is not in any usual sense the return we expect to get. But it is the common name throughout philosophy, economics and statistics, and it is the name I'll use here.)

There's another way to get to calculate expected values. Order each of the possible outcomes from worst to best, and at each step, multiply the probability of getting at least that much by the difference between that amount and the previous step. (At the first step, the 'previous' value is 0.) So Llinos gets £10 with probability 1, has an 0.4 chance of getting another £3, and has an 0.1 chance of getting another £2. Applying the above rule, we work out her expected return is $10 + 0.4 \times 3 + 0.1 \times 2 = 10 + 1.2 + 0.2 = 11.4$. It isn't coincidence that we got

[13] One exception: Robespierre will return from time to time, along with other Terrorists.

[14] Technically, what I'm saying here is that the marginal utility of money to Llinos is constant. There is a usual way of cashing out what it is for the marginal utility of money to be constant in terms of betting behaviour. It is that the marginal utility of money is constant iff the agent is indifferent between a bet that returns $2x$ with probability 0.5, and getting x for sure. But we can't adopt that definition here, because it takes for granted a particular method of valuing bets. And whether that method is correct is about to come into question.

the same result each way; these are just two ways of working out the same sum. But the latter approach makes it easier to understand an alternative approach to decision theory, one recently defended by Lara Buchak (2013).

She thinks that the standard approach, the one I've based around expected values, is appropriate only for agents who are neutral with respect to risk. Agents who are risk seeking, or risk averse, should use slightly different methods.[15] In particular, when we multiplied each possible gain by the probability of getting that gain, Buchak thinks we should instead multiply by some function f of the probability. If the agent is risk averse, then $f(x) < x$. To use one of Buchak's standard examples, a seriously risk averse agent might set $f(x) = x^2$. (Remember that $x \in [0, 1]$, so $x^2 < x$ everywhere except the extremes.) If we assume that this is Llinos's risk function, the bet I described above will have value $10 + 0.4^2 \times 3 + 0.1^2 \times 2 = 10 + 0.48 + 0.02 = 10.5$.

Now imagine a case that is simpler in one respect, and more complicated in another. Iolana has to choose between getting £1 for sure, and getting £3 iff a known to be fair coin lands heads. (The marginal utility of money to Iolana is also constant over the range in question.) And she doesn't know whether she should use standard decision theory, or a version of Buchak's decision theory, with the risk function set at $f(x) = x^2$. Either way, the £1 is worth 1. (I'm assuming that £1 is worth 1 util, expressing values of choices in utils, and not using any abbreviation for these utils.) On standard theory, the bet is worth $0.5 \times 3 = 1.5$. On Buchak's theory, it is worth $0.5^2 \times 3 = 0.75$. So until she knows which decision theory to use, she won't know which option is best to take. That's not merely to say that she won't know which option will return the most. She can't know which option has the best returns until the coin is flipped. It's to say also that she won't know which bet is rational to take, given her knowledge about the setup, until knows which is the right theory of rational decision making.

In the spirit of normative internalism, we might imagine we could solve this problem for Iolana without resolving the dispute between Buchak and her orthodox rivals. Assume that Iolana has, quite rationally, credence 0.5 that Buchak's theory is correct, and credence 0.5 that orthodox theory is correct. (I'm assuming here that a rational agent could have positive credence in Buchak's views. But that's clearly true, since Buchak herself is rational.) Then the bet on the coin has, in some sense, 0.5 chance of being worth 1.5, and 0.5 chance of being worth 0.75. Now we could ask ourselves, is it better to take the £1 for sure, or to take

[15] The orthodox view is that the agent's attitude to risk should be incorporated into their utility function. That's what I think is correct, but Buchak does an excellent job of showing why there are serious reasons to question the orthodoxy.

the bet that has, in some sense, 0.5 chance of being worth 1.5, and 0.5 chance of being worth 0.75?

The problem is that we need a theory of decision to answer that very question. If Iolana takes the bet, she is guaranteed to get a bet worth at least 0.75, and she has, by her lights, an 0.5 chance of getting a bet worth another 0.75. (That 0.75 is the difference between the 1.5 the bet is worth if orthodox theory is true, and the 0.75 it is worth if Buchak's theory is true.) And, by orthodox lights, that is worth $0.75 + 0.5 \times 0.75 = 1.125$. But by Buchak's lights, that is worth $0.75 + 0.5^2 \times 0.75 = 0.9375$. We still don't know whether the bet is worth more or less than the sure £1.

Over the course of this book, we'll see a lot of theorists who argue that in one way or other, we can resolve practical normative questions like the one Iolana faces without actually resolving the hard theoretical issues that make the practical questions difficult. And one common way to think this can be done traces back to an intriguing suggestion by Robert Nozick (1994). Nozick suggested we could use something like the procedure I described in the previous paragraph. Treat making a choice under normative uncertainty as taking a kind of bet, where the odds are the probabilities of each of the relevant normative theories, and the payoffs are the values of the choice given the normative theory.[16] And the point to note so far is that this won't actually be a technique for resolving practical problems without a theory of decision making. At some level, we simply need a theory of decision.

The fully internalist 'theory' turns out to not have anything to say about cases like Iolana's. If it had a theory of second order decision, of how to make decisions when you don't know how to make decisions, it could adjudicate between the cases. But there can't be a theory of how to make decisions when you don't know how to make decisions. Or, more precisely, any such theory will be externalist.

Let's note one striking variant on the case. Wikolia is like Iolana is almost every respect. She gives equal credence to orthodox decision theory and Buchak's alternative, and no credence to any other alternative, and she is facing a choice

[16] Nozick's own application of this was to the Newcomb problem (Nozick 1969). (Going into the details of what the Newcomb problem is would take us too far afield; Paul Weirich (2016) has a nice survey of it if you want more details.) He noted that if causal decision theory is correct, then two–boxing is fractionally better than one–boxing, while if evidential decision theory is correct, then one–boxing is considerably better than two–boxing. If we think the probability that evidential decision theory is correct is positive, and we use this approach, we will end up choosing one box. And that will be true even if the probability we assign to evidential decision theory is very very small.

between £1 for sure, and £3 iff a fair coin lands heads. But she has a third choice: 55 pence for sure, plus another £1.60 iff the coin lands heads. It might be easiest to label her options A, B and C, with A being the sure pound, B being the bet Iolana is considering, and C the new choice. Then her payoffs, given each choice and the outcome of the coin toss, are as follows.

	Heads	Tails
Option A	£1	£1
Option B	£3	£0
Option C	£2.15	£0.55

The expected value of Option C is $0.55 + 0.5 \times 1.6 = 1.35$. (I'm still assuming that £1 is worth 1 util, and expressing values of choices in utils.) It's value on Buchak's theory is $0.55 + 0.5^2 \times 1.6 = 0.95$. Let's add those facts to the table, using EV for expected value, and BV for value according to Buchak's theory.

	Heads	Tails	EV	BV
Option A	£1	£1	1	1
Option B	£3	£0	1.50	0.75
Option C	£2.15	£0.55	1.35	0.95

Now rememeber that Wikolia is unsure which of these decision theories to use, and gives each of them equal credence. And, as above, whether we use orthodox theory or Buchak's alternative at this second level affects how we might incorporate this fact into an evaluation of the options. So let EV2 be the expected value of each option if it is construed as a bet with an 0.5 chance of returning its expected value, and an 0.5 chance of returning its value on Buchak's theory, and BV2 the value of that same bet on Buchak's theory.

	Heads	Tails	EV	BV	EV2	BV2
Option A	£1	£1	1	1	1	1
Option B	£3	£0	1.50	0.75	1.125	0.9375
Option C	£2.15	£0.55	1.35	0.95	1.15	1.05

And now something interesting happens. In each of the last two columns, Op-

tion C ranks highest. So arguably[17], Wikolia can reason as follows: *Whichever theory I use at the second order, option C is best. So I should take option C*. On the other hand, Wikolia can also reason as follows. If expected value theory is correct, then I should take option B, and not take option C. And if Buchak's theory is correct, then I should take option A, and not take option C. So either way, I should not take option C. Wikolia both should and should not take option C.

That doesn't look good, but again I don't want to overstate the difficulty for the internalist. The puzzle isn't that internalism leads to a contradiction, as it might seem here. After all, the term 'should' is so slippery that we might suspect there is some kind of fallacy of equivocation going on. And so our conclusion is not really a contradiction. It really means that Wikolia should-in-some-sense take option C, and should-not-in-some-other-sense take option C. And that's not a contradiction. But it does require some finesse for the internalist to say just what these senses are. This kind of challenge for the internalist, the puzzle of ending up with more senses of should than one would have hoped for, and needing to explain each of them, will recur a few times in the book.

1.7. Two Recent Debates

I think the question of whether *Do the right thing* or *Follow your principles* is more fundamental is itself an interesting question. But it has become relevant to two other debates that have become prominent in recent philosophy as well. These are debates about moral uncertainty, and about higher-order evidence.

Many of the philosophers who have worried that *Do the right thing* is insufficiently guiding have looked to have a theory that makes moral uncertainty more like factual uncertainty. And since it is commonly agreed that an agent facing factual uncertainty, and only concerned with outcomes, should maximise factual uncertainty, a common conclusion has been that a morally uncertain agent should also maximise some kind of expected value. In particular, they should aim to maximise the expected moral value of their action, where probabilities about moral theories can affect the expected moral value.

[17] Ironically, it isn't at all obvious in this context that this is acceptable reasoning on Wikolia's part. The argument by cases she goes on to give is not strictly speaking valid on Buchak's theory, so it isn't clear that Wikolia can treat it as valid here, given that she isn't sure which decision theory to use. This goes to the difficulty of saying anything about what should be done without making substantive normative assumptions, a difficulty that will recur frequently in this book.

In the recent literature, we see the view that people should be sensitive to the probabilities of moral theories sometimes described as 'moral hedging'. This terminology is used by Christian Tarsney (2017), who is fairly supportive of the idea, and Ittay Nissan-Rozen (2015), who is not. It's not, I think, the happiest term. After all, Robespierre maximised expected moral value, at least relative to the credences that he had. And it would be very odd to describe the Reign of Terror as a kind of moral hedging.

The disputes about moral uncertainty have largely focussed on cases where a person is torn between two (plausible) moral theories, and has to choose between a pair of actions. In one important kind of case, the first is probably marginally better, but it might be much much worse. In that case, maximising moral value may well involve taking the second option. And that's the kind of case where it seems right to describe the view as a kind of moral hedging.

But the general principle that one should maximise expected moral value applies in many more cases than that. It applies, for example, to people who are completely convinced that some fairly extreme moral theory is correct. And in those cases, maximising expected moral value, rather than actual moral value, could well be disastrous.

When it is proposed that probabilities matter to a certain kind of decision, it is a useful methodology to ask what the proposal says in the cases where the probabilities are all 1 or 0. That's what I'm doing here. If probabilities of moral theories matter, they should still matter when the probability (in the relevant sense) of some horrid theory is 1. So my investigation of Polonius's principle will have relevance for the debate over moral uncertainty, since it will have consequences for what theories of moral uncertainty can plausibly say in extreme cases.

There is one dispute about moral uncertainty that crucially involves intermediate probabilities. Maximising expected moral value requires putting different theories' moral evaluations of actions on a common scale. There is no particularly good way to do this, and it has been argued that there is no possible way to do this. This is sometimes held to be a reason to reject 'moral hedging' (Hedden 2016a). I'll return to this question in chapter 6, offering a tentative defence of the 'hedger'. The question of how to find this common scale is hard, but there are reasons to think it is not impossible. And what matters for the current debate is whether it is in fact impossible.

The other recent dispute that normative externalism bears on concerns peer disagreement. Imagine that two friends, Ankita and Bojan, both regard themselves and each other as fairly knowledgable about a certain subject matter. And let p

be a proposition in that subject, that they know they have equally good evidence about, and that they are antecedentl equally likely to form true beliefs about. Then it turns out that Ankita believes p, while Bojan believes $\neg p$. What should they do in response to this news?

One response goes via beliefs about their own rationality. Each of them should think it is equally likely that believing p and believing $\neg p$ is rational given their common evidence. They should think this because they have two examples of rational people, and they ended up with these two conclusions. So they should think that holding on to their current belief is at most half-likely to be rational. And it is irrational, say some theorists, to believe things that you only think are half-likely to be rational. So both of them should become completely uncertain about whether p is true.

I'm going to argue that there are several mistakes in this reasoning. They shouldn't always think that holding on to their current belief is half-likely to be rational. Whether they should or not depends, among other things, on why they have their current belief. But even if they should change their belief about how likely it is that their belief is rational, nothing follows about what they should do to their first-order beliefs. In some strange situations, the thing to do is to hold on to a belief, while being sceptical that it is the right belief to have. This is the key externalist insight, and it helps us resolve several puzzles about disagreement.

1.8. Elizabeth and Descartes

Although the name 'normative externalism' is new, the view is not. It will be obvious in what follows how much the arguments I have to offer are indebted to earlier work by, among others, Nomy Arpaly (2003), Timothy Schroeder (Arpaly and Schroeder 2014), Maria Lasonen-Aarnio (2010a, 2014a), Miriam Schoenfield (2015) and Elizabeth Harman (2011, 2015). It might not be as obvious, because they aren't directly cited as much, but much of the book is influenced by the pictures of normativity developed by Thomas Kelly (2005) and by Amia Srinivasan (2015b).

Many of the works just cited address just one of the two families of debates this book joins: i.e., debates about ethics and debates about epistemology. One of the nice features about taking on both of these debates at once is that it is possible to blend insights from the externalist side of each of those debates. So chapter 4,

which is the main argument against normative internalism in ethics, is modelled on an argument Miriam Schoenfield (2015) develops to make a point in epistemology. And much of what I say epistemic akrasia in chapter 10 is modelled on what Nomy Arpaly (2003) says about practical akrasia.

There are also some interesting historical references to normative externalism. I'm just going to talk about the one that is most interesting to me. In the correspondence between Descartes and Elizabeth, we see Descartes taking a surprisingly internalist view in ethics, and Elizabeth the correct externalist view.[18]

On 15 September, 1645, Descartes wrote:

> For it is irresolution alone that causes regret and repentance.

This had been a theme of the view he had been putting forward. The good person, according to the view Descartes put forward in the correspondence, is one who makes a good faith effort to do the best they can. Someone who does this, and who is not irresolute, has no cause to regret their actions. He makes this clear in an earlier letter, on 4 August 1645, where he is also more explicit that it is only careful and resolute actors who are immune to regret.

> But if we always do all that our reason tells us, we will never have any grounds to repent, even though events afterward make us see that we were mistaken. For our being mistaken is not our fault at all.

Elizabeth disagrees with Descartes both about regret, and with what it shows us about the nature of virtue. She writes, on 16 August 1645,

> On these occasions regret seems to me inevitable, and the knowledge that to err is as natural to man as it is to be sick cannot protect us. For we also are not unaware that we were able to exempt ourselves of each particular fault.

Over the course of the correspondence, Elizabeth seems to be promoting a view of virtue on which being sensible in forming intentions, and resolute in carrying them out, does not suffice for being good. One must also form the right intentions. If that is really her view, then she is a very important figure in the history of normative externalism. Indeed, if that is her view, perhaps I should be calling this book a defence of Elizabethan philosophy.

[18] All translations are from the recent edition of the correspondence by Lisa Shapiro (Elizabeth and Descartes 2007).

But it would be a major diversion from the themes of this book to investigate exactly how much credit Elizabeth is due. And in any case, I don't want to suggest that I'm defending exactly the view Elizabeth is. The point about possibility she makes in the above quote is very important. It's possible that we ought to be good, and we can't know just what is good, but this isn't a violation of *Ought implies can*, because for any particular good thing we ought do, we can with effort come to know that that thing is good. That's a nice problem to raise for particular internalists, but it's not my motivation for being externalist. I don't think it matters at all whether we know what is good, so the picture of virtue I'm working with is very different to the Stoic picture that Elizabeth has. (It's much more like the picture that Nomy Arpaly (2003) has developed.)

So it would be misleading to simply say this book is a work in Elizabethan philosophy. But Elizabeth is at the very least an important figure in the history of the views I'm defending, and she is to me the most fascinating of my historical predecessors.

1.9. Why Call This Externalism?

There are so many views already called externalist in the literature that I feel I should offer a few words of defence of my labelling my view externalist. In the existing literature I'm not sure there is any term, let alone an agreed upon term, for the view that higher-order considerations are irrelevant to both ethical and epistemological evaluation. So we needed some nice term for my view. And using 'externalist' suggested a useful term for the opposing view. And there is something evocative about the idea that what's distinctive of my view is that it says that agents are answerable to standards that are genuinely *external* to them. More than that, it will turn out that there are similarities between the debates we'll see here and familiar debates between internalists and externalists about both content and about the nature of epistemic norms.

In debates about content, we should not construe the internalism/externalism debate as a debate about which of two kinds of content are, as a matter of fact, associated with our thought and talk. To set up the debate that way is to concede something that is at issue in the debate. That is, it assumes from the start that there is an internalist friendly notion of content, and it really is a kind of content. But this is part of what's at issue. The same point is true here. I very much do not think the debate looks like this: The externalist identifies some norms, and the internalist identifies some others, and then we debate which of those norms

are really our norms. At least against some internalist opponents, I deny that they have so much as identified a kind of norm that we can debate whether it is our norm.

In debates in epistemology, there is a running concern that internalist norms are really not normative. If we identify justified belief in a way that makes it as independent of truth as the internalist wants justification to be, there is a danger that that we should not care about justification. Internalists have had interesting things to say about this danger (Conee 1992), and I don't want to say that that it is a compelling objection to (first-order epistemological) internalism. But it is a danger. And I will argue that it's a danger that the normative internalist can't avoid.

Let's say we can make sense of a notion that tracks what the internalist thinks is important. In section 6.1 I'll argue that *not being a hypocrite* is such a notion; the internalist cares a lot about it, and it is a coherent notion. There is a further question of whether this should be relevant to our belief, our action or our evaluation of others. If someone is a knave, need we care further about whether they are a sincere or hypocritical knave? I'll argue that at the end of the day we should not care; it isn't worse to be hypocritical.[19]

The debates I'm joining here have something else in common with familiar internalist/externalist debates. Many philosophers will be tempted to react to them by saying the parties are talking at cross-purposes. In fact, there are two ways that it might be thought the parties are at cross-purposes.

First, it might be thought the parties are both right, but they are talking about different things. The normative internalist is talking about subjective normativity, and saying true things about it, while the normative externalist is talking about objective normativity, and saying true things about it. One of the running themes of this book will be that this isn't a way of dissolving the debate, it is a way of taking the internalist's side. Just like in debates about content, and in debates about epistemic justification, the externalist denies that there is any notion that plays the role the internalist wants their notion to play. To say the notion exists, but isn't quite as important as the internalist says it is, is to concede the vast majority of what the externalist wants to contest.

[19] My instinct is that there is something preferable about the hypocrite compared to the person who does wrong while thinking they are doing the right thing. After all, the hypocrite has figured out a moral truth, and figuring out moral truths typically reflects well on a person. But I'm not going to try to turn this instinct into an argument in this book.

The second way to say that the theorists are talking at cross-purposes is to say that their differences merely turn on first-order questions about ethics and epistemology. What the internalist calls misleading evidence about morality, the externalist calls first-order reasons to act a different way. And what the internalist calls higher-order evidence, the externalist calls just more first-order evidence. This is, I'm going to argue, an externalist position, and not one that the internalist should happily sign on to. It is, very roughly, the view I want to defend in epistemology. What has been called higher-order evidence in epistemology is, when it is anything significant at all, just more first-order evidence. It is also a possible externalist view in ethics, though not one I want to defend. In particular, it is the view that misleading evidence about morality changes the objective circumstances in a way that changes what is good to do. I don't think that's typically true, but it is a possible externalist view.

All that said, there are two ways in which what I'm saying differs from familiar internalist/externalist debates. One is that what I'm saying cross-cuts the existing debates within ethics and epistemology that often employ those terms. Normative externalism is compatible with an internalist theory of epistemic justification. It is consistent to hold the following two views:

- Whether S is justified in believing *p* depends solely on S's internal states.
- There is a function from states of an agent to permissible beliefs, and whether an agent's beliefs are justified depends solely on the nature of that function, and the agent could in principle be mistaken, and even rationally mistaken, about the nature of the function.

The first bullet point defines a kind of internalism in epistemology. The second bullet point defines a kind of externalism about epistemic norms. But the two bullet points are compatible, as long as the function in question does not vary between agents with the same internal states. The two bullet points may appear to be in some tension, but their conjunction is more plausible than many theses that have wide philosophical acceptance. Ralph Wedgwood (2012), for example, defends the conjunction, and spends some time arguing against the idea that the conjuncts are in tension.

And normative externalism is compatible in principle with the view in ethics that there is an internal connection between judging that something is right, and being motivated to do it. This view is sometimes called motivational internalism (Rosati 2016). But again, there is a tension, in this case so great that it is hard to see why one would be a normative externalist and a motivational internalist. The tension is that to hold on to both normative externalism and motivational

internalism simultaneously, one has to think that 'rational' is not an evaluative term, in the sense relevant for the definition of normative externalism. That is, one has to hold on to the following views.

- It is irrational to believe that one is required to φ, and not be motivated to φ; that's what motivational internalism says.
- An epistemically good agent will follow their evidence, so if they have misleading moral evidence, they will believe that φ is required, even when it is not. The possibility of misleading moral evidence is a background assumption of the debate between normative internalists and normative externalists. And the normative externalist says that the right response to misleading evidence is to be misled.
- An agent should be evaluated by whether they do, and are motivated to do, what is required of them, not whether they do, or are motivated to do, what they believe is required of them. Again, this is just what normative externalism says.

Those three points are consistent, but they entail that judging someone to be irrational is not, in the relevant sense, to evaluate them. Now that's not a literally incoherent view. It is a souped-up version of what Niko Kolodny (2005) argues for. (It isn't Kolodny's own view; he thinks standards of rationality are evaluative but not normative. I'm discussing the view that they are neither evaluative nor normative.) But it is a little hard to see the attraction of the view. So normative externalism goes more happily with motivational externalism.

And that's the common pattern. Normative externalism is a genuinely novel kind of externalism, in that it is neither entailed by, nor entails, other forms of externalism. But some of the considerations for and against it parallel considerations for and against other forms of externalism. And it sits most comfortably with other forms of externalism. So the name is a happy one.

1.10. Plan of Book

This book is in two parts: one about ethics, the other about epistemology.

The ethics part starts with a discussion of the motivations of internalism in ethics. It then spends two chapters arguing against strong forms of internalism. By strong forms, I mean views where some key moral concept is identified with acting in accord with one's own moral beliefs. So this internalist-friendly condition (I'm doing what I think I should do) is both necessary and sufficient for

some moral concept to apply. After this, I spend two chapters on weak forms. In chapter 5, I discuss a view where blameworthiness requires that one not believe one was doing the wrong thing. In chapter 6, I discuss a view where doing what one thinks is wrong manifests a vice, even if the action is right. These don't cover the field of possible views, but they are important versions of views that hold that internalist-friendly conditions have a one-way connection to key moral concepts. The internalist-friendly conditions in these cases provide either a necessary or a sufficient condition for the application of a key moral concept, but not both.

I then turn to epistemology. The organising principle that I'll be defending is something I'll call Change Evidentialism: only new evidence that bears on p can compel a rational agent to change their credences in p. The forms of internalism that I'll be opposing all turn out to reject that. And the reason they reject it is that they think a rational person can be compelled to change their credences for much more indirect reasons. In particular, the rational person could get misleading evidence that the rational attitude to take towards p is different to the attitude they currently take, and that could compel them to change their attitude towards p. I'm going to argue that this is systematically mistaken. And this has consequences for how to think about circular reasoning (it's not as bad as you think!), epistemic akrasia (it's not as bad as you think!), and standing one's ground in the face of peer disagreement (it's really not as bad as you think!).

Part I.

Ethics

2. All About Internalism

This chapter has two related aims. The first is to clarify, and classify, the range of internalist positions that are available. The second is to set out more carefully the reasons for adopting one or other of these positions. We'll end by putting the two parts together, seeing which motivations push towards what kinds of internalism. These themes were all introduced briefly in the introduction, but they need fuller treatment before we proceed.

It is always good practice to state as carefully and as persuasively as possible the view one means to oppose. But there is a particular reason to adopt that general practice here. Some of the appeal of internalism comes from sliding between different versions of the view. Once we get some key distinctions on the table, we get a better look at which versions are defensible.

The conclusion of the chapter will be that the best arguments for normative internalism in ethics make heavy use of the idea that moral uncertainty and factual uncertainty should be treated symmetrically. So to get started, we'll look at how factual uncertainty matters morally.

2.1. Some Distinctions

It helps to have some mildly technical language on the table to begin with. The terminology I'll use here is standard enough. But the terms are somewhat ambiguous, and theoretically loaded. I want to stipulate away some possible ambiguities, and simultaneously avoid at least some theoretical disputes. So take the following elucidations of the distinctions to be definitional of the bolded terms as they'll be used here.

- Useful vs Harmful *Outcomes.* Some outcomes involve more welfare, others involve less. I'll say an action is more useful to the extent that it in-

volves more welfare, and harmful to the extent it involves less.[1]

- Good vs Bad *Outcomes*. Some outcomes are better, all things considered, than others. I'll use good and bad as predicates of outcomes, ones that track whether the outcome is better or worse. It is common enough to talk about good and bad actions, and good and bad agents, but I'll treat those usages as derivative. What's primary is whether outcomes are good or bad. I will *not* assume that the goodness of an outcome is agent-independent. Perhaps an outcome where a person lies to prevent a great harm is bad relative to that person, since they have violated a categorical moral imperative. That is consistent with saying the lie was very useful, and even that it was good relative to other people.
- Right vs Wrong *Actions*. Unlike *good* and *bad*, I'll use *right* and *wrong* exclusively as predicates of actions.
- Rational vs Irrational *Actions* and *States*. This is a bit of a stretch of ordinary usage, but I'll talk both about mental states (beliefs, intentions, etc.) being rational or irrational, and the actions that issue from these states being rational or irrational. So it is both irrational to believe that the moon is made of green cheese, and to bet that it is.
- Praiseworthy vs Blameworthy *Agents*. Again, there is an ordinary usage where actions are praiseworthy or blameworthy. But I'll treat that as derivative. What's primary is that an agent is praiseworthy or blameworthy, perhaps in virtue of having performed a particular action.

In conditions of full knowledge, it is very plausible that there are close connections between these five distinctions. There is a natural form of consequentialism that says the five are co-extensive under conditions of full knowledge. A good outcome just is a useful one; a right action is one that promotes the good; it is rational to promote the good, and blameworthy to do not so. Those who are not sympathetic to classical consequentialism will not be happy with this equation between the good and the useful, but they might support many of the other equations. Michael Smith (2006, 2009) for example, has argued that if we allow goodness to be agent-relative, then even non-consequentialists can allow that, under conditions of full knowledge, right actions are those that maximise the good. Smith's argument is not uncontroversial. Campbell Brown (2011) notes there will be problems with this attempt to 'consequentialize' a theory that allows for moral dilemmas. But I'm going to set that issue aside.

Under conditions of uncertainty, the connections between the distinctions be-

[1] I'm going to stay neutral about just what outcomes are. I prefer to think of them as possible worlds, but there are many other choices that would do just as well for current purposes.

comes much more murky, even for a consequentialist. There are cases where the useful comes apart from the right, the rational, and the praiseworthy. Here are two such cases.

> Cressida is going to visit her grandmother, who is unwell, and who would like a visit from her granddaughter. She knows the more time she spends with her grandmother, the better things will be. So she drives as fast as she can to get there, not worrying about traffic lights or any other kind of traffic regulation. Normally this kind of driving would lead to several serious injuries, and possibly to fatalities, but by sheer good fortune, no one is harmed by Cressida's driving. And her grandmother does get some enjoyment from spending a few more minutes with her granddaughter.

> Botum is the chief executive of a good, well-run, charity. She has just been given a £10,000 donation, in cash. She is walking home her normal way, through the casino. As she is walking past the roulette table, it occurs to her that if she put the £10,000 on the right number, she could turn it into £360,000, which would do much more good for the charity. She has 38 choices: Do nothing, bet on 0, bet on 1, ..., bet on 36. Of these, she knows the one with the most useful outcome will be one of the last 37. But she keeps the money in her pocket, and deposits it with the charity's bank account the next morning.

Cressida acts wrongly, and is seriously blameworthy for her driving. That's even though the outcome is the best possible outcome. So there's no simple connection, given uncertainty, between usefulness and rightness.

But in some ways the case of Cressida is simple. After all, it is very improbable that driving this way will be useful. We might think that there is still a duty to maximise the probability of being maximally useful. The case of Botum shows this isn't true. She does the one thing she knows cannot be maximally useful. But that one thing is the one and only right thing for her to do. All the other alternatives are both wrong and blameworthy, and that includes the one very useful one.

This way of talking about right and wrong is not universally adopted. In part this is an unimportant matter of terminological regimentation, but I suspect in part it reflects a deeper disagreement. Here's the kind of case that motivates the way of talking I'm not going to use.

> Adelajda is a doctor, and Francesc her patient. Francesc is in a lot of pain, so Adelajda provides pain medication to Francesc. Unfortunately, someone wants to kill Francesc, so the pain medication has been adulterated. In fact, when Adelajda gives Francesc this medicine, she kills him.

A common verdict on this kind of case is that Adelajda acts wrongly, since she kills someone, but blamelessly, since she was ignorant of what she was injecting Francesc with (Rosen 2008; Graham 2014; E. Harman 2015). The picture seems to be that an action is wrong if it brings about a bad outcome, and considerations of what was known are irrelevant to the wrongnes of the act. So Adelajda's act is wrong because it is a killing, independent of her knowledge.

I think this is at best an unhelpful way to think about Adelajda. In any case, I'm not going to use 'right' and 'wrong' in that way. On my preferred picture, Adelajda's ignorance doesn't provide her an excuse, because she didn't do anything wrong. (I follow orthodoxy in thinking that excuses are what make wrong actions less blameworthy.) I think the picture where Adelajda doesn't do anything wrong makes best sense of cases like Botum's. I'm here following Frank Jackson (1991), who supports this conclusion with a case like this one.

> Billie is a doctor, and Jack her patient. Jack has a very serious disease. He is suffering severe stomach pains, and the disease will soon kill him if untreated. There are three drugs that would cure the disease, A, B and C. One of A and B would stop Jack's pain immediately, and cure the disease with no side effects. The other would have side effects so severe they would kill Jack. Billie has no idea which is which, and it would take two days of tests to figure out which to use, during which time Jack would suffer greatly. Drug C would cure the disease, but cause Jack to have one day of severe headaches, which would be just as painful as the stomach pains he now has.

The thing for Billie to do is to give Jack drug C. (I'm saying 'thing to do' rather than using a term like 'good or 'right' because what's at issue is figuring out what's good and right.) Giving Jack drug A or B would be a horribly reckless act. Waiting to find out which of them would have no side effect would needlessly prolong Jack's suffering. So the thing to do is give him drug C.

But now consider things from the perspective of someone with full knowledge. (Maybe we could call that the objective perspective, but I suspect the terminology

of 'objective' and 'subjective' obscures more than it reveals here.) Billie directly causes Jack to have severe headaches for a day. This was avoidable; there was a drug that would have cured the disease with no side effects at all. Given full knowledge, we can see that Billie caused someone in her care severe pain, when this wasn't needed to bring about the desired result. This seems very bad.

And things get worse. We can imagine Billie knows everything I've said so far about A, B and C. So she knows, or at least could easily figure out, that providing drug C would be the wrong thing to do if she had full knowledge. So unlike Adelajda, we can't use her ignorance as an excuse. She is ignorant of something all right, namely whether A or B is the right drug to use. But she isn't ignorant of the fact that providing C is wrong given full information. Now assume that we should say what Adelajda does is wrong (since harmful), but excusable (because she does not and could not know it is wrong). It follows that what Billie does is also wrong (since harmful) but not excused (since she does know it is wrong).

This all feels like a reductio of that picture of wrongness and excuse. The full knowledge perspective, independent of all considerations about individual ignorance, is not constitutive of right or wrong. Something can be the right thing to do even if one knows it will produce a sub-optimal outcome. So it can't be ignorance of the effects of one action provides an excuse which makes a wrong action blameless. Billie needs no excuse, even though she needlessly causes Jack pain. That's because Billie does nothing wrong in providing drug C. Similarly, Adelajda does nothing wrong in providing the pain medication. In both cases the outcome is unfortunate, extremely unfortunate in Adelajda's case. But this doesn't show that their actions need excusing, and doesn't show that what they are doing is wrong.

The natural solution here is to say that what is right for Botum or Billie to do is not to maximise the probability of a useful outcome, but to maximise something like expected utility. It won't matter for current purposes whether we think Botum should maximise expected utility itself, or some other risk-adjusted value, along the lines suggested by John Quiggin (1982) or Lara Buchak (2013). The point is, we can come up with a 'subjective' version of usefulness, and this should not be identified with the probability of being useful. We'll call cases like Botum and Billie's, where what's right comes apart from even the probability of being best, Jackson cases, and return to them frequently in what follows.[2]

[2]Similar cases were discussed by Donald Regan (1980) and Derek Parfit (1984). But I'm using the terminology 'Jackson case' since my use of the cases most closely resembles Jackson's, and because the term 'Jackson case' is already in the literature.

Expected values are only defined relative to a probability function. So when we ask which action maximises expected value, the question only has a clear answer if we make clear which probability functions we are talking about. Two probability functions in particular will be relevant going forward. One is the 'subjective' probability defined by the agent's credences. The other is the 'evidential' probability that tracks how strongly the agent's evidence supports one proposition or another. These will generate subjective expected values, and evidential expected values, for each possible action. And both values will have a role to play in later discussion.

2.2. Two Ways of Maximising Expected Goodness

So far we have only looked at agents who are uncertain about a factual question. Cressida does not know who she will harm by driving as she does, Botum does not know which number will come up on the roulette wheel, and Adelajda and Billie are ignorant of the effects of some medication. But we could also imagine that agents are uncertain about normative questions.

Deorsa is deciding whether to have steak or tofu for dinner. He is a remarkably well informed eater, and so he knows a lot about the process that goes into producing a steak. But try as he might, he can't form an opinion on the moral appropriateness of eating meat. He thinks meat eating results in outcomes that are probably not bad, but like many carnivores, he has his doubts.

To simplify the story, I'm going to make three assumptions. The first assumption is that Deòrsa is actually in a world where meat eating is not bad. The second assumption is that Deòrsa is perfectly reasonable in having a high, but not maximal, credence in meat eating not being bad. You may think that this requires Deòrsa to live in a world very unlike this one, or even an impossible world. But that's OK for the story I'm telling; I just need Deòrsa's situation to be conceivable. (We will spend a lot of time thinking about impossible worlds as this book goes on, so it's useful to warm up with one that might be impossible now.) And the third assumption is that there is a large asymmetry between Deòrsa's choices. If meat eating is not bad, it would be ever so slightly better for Deòrsa to have the steak, since he would get some enjoyment from it, and it wouldn't be bad in any other respect. But if meat eating is bad, then having the steak would be a much much worse outcome, since it would involve Deòrsa in an unjustified killing.

Which action, having the steak or having the tofu maximises expected goodness? That question is ambiguous. In one sense the answer is tofu. After all, there is a non-trivial probability that having the steak leads to a disasterous outcome. In another sense, the answer is steak. After all, there is a thing goodness, and Deòrsa knows enough to know of it that it is maximised by steak eating. Since Deòrsa is to some extent morally ignorant, he doesn't know what goodness is, so he thinks goodness might be something else, something that is not maximised by steak eating. But given his (perfectly reasonable, rational) credences, the thing that is goodness has its expected (and actual) value maximised by steak eating.

We might put the distinction in the previous paragraph by saying that the action that maximises the expected value of goodness de re, that is, of the thing that is goodness, is different from the action that maximises the expected value of goodness de dicto, that is, of whatever it is that goodness turns out to be. And using the de dicto/de re terminology, we can see that this distinction applies across a lot of realms. Here are two more examples where we can use it.

> Monserrat is playing the board game Settlers of Catan. She has to decide between two moves. She is uncertain how the moves will affect the later game play. This is reasonable, since the game play includes dice rolls that she couldn't possibly predict. But she's also forgotten what the victory condition is. She can't remember if it is first to 10 points wins, or first to 12 points. The standard is 10, but some games are played under special house rules that change this. In Monserrat's game, there aren't any special house rules, so it is actually 10 points that wins. Call the moves that she is choosing between A and B. If she plays A, she has a 30% chance of being first to 10 points, and a 50% chance of being first to 12 points. If she plays B, she has a 40% chance of being first to 10 points, but only a 10% chance of being first to 12 points. She thinks it is 60% likely that the winner is the first to 10, and 40% likely that the winner is the first to 12. So playing B maximises the probability of winning de re. That is, it maximises the probability of doing the thing that is actually winning, i.e., being first to 10. But playing A maximises the probability of winning de dicto. Given Monserrat's uncertainty about the victory conditions, she thinks her probability of winning is 38% if she plays A, and only 34% if she plays B.
>
> A professor is deciding which music to put on. She would prefer lowbrow, trashy music. But, suffering from a common enough

> kind of false consciousness, she thinks she would prefer highbrow, classy music. So playing the lowbrow music would maximise expected preference satisfaction de re. That is, it would maximise the expected value of the satisfaction level of the preferences she actually has. But playing the classy music would maximise expected preference satisfaction de dicto. That is, given her beliefs about her preferences, it seems that the classy music would do a better job at satisfying her preferences.

The key internalist idea is that in situations that call for maximising expected goodness (or utility, or anything else), it is the de dicto version, not the de re version, that matters. The key externalist idea is that it is the de re version that matters. For the rest of this chapter, while I'm setting up and motivating internalism, I'll leave it tacit that we are talking about expected values de dicto.

2.3. Varieties of Internalism

The chapter started with a five-way distinction between the useful, the good, the right, the rational and the praiseworthy. And we noted that for each of those, there were three separate questions we can ask in any practical situation. First, we can ask what action would be most useful/good/right/rational/praiseworthy. Second, we can ask what action has the highest expected usefulness/goodness/rightness/rationality/praiseworthiness given the credences of the agent. Third, we can ask that same question, but relativise the answer to the agent's evidence, not the agent's credences. Multiplying the five way distinction by the three types of question gives us fifteen questions. And each of those fifteen questions picks out a kind of standard. It is an interesting feature of a possible choice that it actually is the rational one, or that it maximises credal expected praiseworthiness, or evidential expected usefulness. For now, call the questions about what actually is most useful etc objective questions, and the standard that an action or choice meets in virtue of being the answer to such a question an objective standard. (This is just to distinguish the first class of questions from the credal and evidential questions.)

Having these fifteen standards in mind, the five objective standards, the five credal standards and the five evidential standards, we have the resources to formulate a number of interesting internalist theses. The theses I have in mind are of the form:

- X objectively meets normative standard N1 when she meets credal/evidential standard N2.

Philosophers who endorse these theses usually take it that the explanatory direction here goes from right-to-left. It is because the agent meets credal/evidential standard N2 that she objectively meets standard N1. But my primary focus will be on the truth of these claims, and not yet the claims about explanatory priority.

Michael Zimmerman (2008) endorses the following two theses, which exemplify this schema.

- An action is right when it maximises evidentially expected goodness, and it is wrong when it does not.
- A person is praiseworthy for maximising credally expected goodness, and blameworthy for not doing so.

Michael Smith (2006, 2009) argues (against the arguments from Jackson I gave above) that right action is just action that maximises the good. But what an agent is responsible for is whether they maximise evidential expected goodness de dicto. Indeed, what they should do, in 'the sense most relevant for action', is maximise evidential expected goodness de dicto (M. Smith 2006, 144). Moreover, this is what rationality requires (M. Smith 2009).

There are obviously a lot of other possibilities for N1 and N2 that we could use, and that gives us a lot of internalist theses. Before we go on, three clarifications on what I am, and what I am not, counting as an internalist thesis.

First, I've put the statements above in ways that are naturally interpreted as universal quantifications. That makes them very strong, perhaps implausibly strong. A view that said that theses like the above held ceteris paribus, or held subject to side constraints, or held in a well defined range of cases, would still be internalist in the sense I'm interested in.

Second, the theses listed above are biconditionals. We could weaken them to one-way conditionals, and still get something recognisably internalist, as long as we think that the conditional is still somewhat explanatory. For instance, a view that said an agent is blameless for what they do if they maximise evidential expected goodness would be internalist, even if it didn't give necessary and sufficient conditions for blamelessness. Such a view might also add some externalist conditions to blamelessness; perhaps it would go on to say that someone

is blameless as long as they actually don't make things worse, or actually do anything wrong. It's a matter of terminological preference whether we count these hybrid views as internalist or externalist, but since I plan to argue against them, I'm counting them as internalist. (Chapters 5 and 6 will be dedicated to a discussion of some such views.)

Third, I'm not counting a view as internalist unless both N1 and N2 are person-evaluative. What I mean by saying a term is person-evaluative is that it is a term we use for evaluations that essentially apply to persons, or actions or states of persons. So truth is not person-evaluative, since we can ask whether the output of a measuring device is true, and harmfulness is not person-evaluative, since earthquakes and volcanoes are harmful. But rationality, praiseworthiness, moral goodness, and moral rightness are person-evaluative (at least if they are evaluative).

So the view Jackson (1991) defends, where rightness is a matter of maximising expected benefits, is not internalist in my sense, because being a benefit is not a person-evaluative notion. Put another way, we don't positively evaluate Cressida the reckless driver, even if we note that her actions actually had a small benefit to the world.

A harder case to judge is whether this should count as an internalist thesis.

- It is a requirement of rationality that one does the thing that maximises expected goodness (de dicto).

Is that an internalist thesis, or not? It depends on what one thinks about rationality. Is rationality person-evaluative. Well, it essentially applies to people. (If we judge a machine is thinking rationally, and not just accurately, we are treating it as a person.) But is it evaluative? It's easy to think this question is easy. Ideal agents are rational, and it is good to be like ideal agents, so of course it is good to be rational. But that's too quick. An ideal taker of a logic quiz would make an even number of errors, since they would make 0 errors, and 0 is even. But that doesn't mean the property of making an even number of errors is an evaluative notion in any sense. We shouldn't say, "Good for you, you made an even number of errors." Making an even number of errors seems completely epiphenomenal from an evaluative standpoint. And it would be an absurd thing to aim at, as such. It's surprising how common it is that properties of the ideal are actually bad to aim at, since they often make things worse in the absence of other features of the ideal (Lipsey and Lancaster 1956-1957). If one thinks being rational is like possessing the property *makes an even number of mistakes*, then one could

agree that rationality involves maximising expected goodness, without thereby disagreeing with externalism.

Now as a matter of fact, I personally think rationality is evaluative, and is not a matter of maximising expected goodness. So I think the thesis is internalist, and is false. But the classificatory question is still important. After all, this thesis is certainly true:

- An action maximises expected goodness iff it maximises expected goodness.

This looks like it has the structure of my canonical internalist theses, with N1 being *maximises expected goodness* and N2 being *goodness*. So doesn't this show that some internalist theses are true? No, I say. This isn't internalist because maximising expected goodness, where this is understood de dicto and not de re, is not a positive feature of a person. It is a feature that ideal agents have, but it is also a feature that political fanatics like Robespierre have. And it isn't a good-making feature in either of them. Rather, it is like making an even number of errors; something that can be instantiated in very good ways, or very bad ways.

2.4. An Initial Constraint

The internalist schema above has some interesting instances when N1 = N2. For instance, we could consider the following theories, where we use the same kind of evaluation on both sides of the biconditional.

- It is right to maximise the expected rightness of one's actions, and wrong to do otherwise.
- It is blameworthy to do what is most probably blameworthy.

But there is a quick argument that all such principles are mistaken. The brief version of the argument is that no such principle is compatible with the conjunction of knowledge of one's own mental states, plus uncertainty about what I'll call morally asymmetric choices. But there is nothing wrong with knowing one's own mental states when faced with a morally asymmetric choice, so the principles must be wrong.

A morally asymmetric choice is where we know that one side of the choice is not in any way morally problematic. A simple case, for most people, is the choice between meat eating and vegetarianism. Very few people would think that it is

immoral, bad, wrong, or blameworthy to be vegetarian on ethical grounds. On the other hand, it is easy to feel some qualms about eating meat. So it looks like this is a choice where all the moral risk falls on one side.

(I'm more interested in the general principle than the particular case, but let me note two quick complications before moving on. It's imaginable that there is a person who puts either their own health or, if they are pregnant or nursing, their child's health, at risk by not eating any meat. In the situations most readers of this book find themselves, those situations will be vanishingly rare since there are so many meat alternatives available. But it's at least conceivable. In the cases I'm discussing I want it to be explicitly part of the case that the person making the choice faces no health complications from being vegetarian. Second, I'm ignoring the possibility that denying oneself pleasures for spurious reasons is immoral. It would merely complicate, but not overturn, the argument to allow for that possibility.)

Now let's think about the first bulleted principle above, which I'll call ProbWrong. And consider an agent who is deciding between steak and tofu for dinner. Imagine that she has the following mental states:

1. She is sure that ProbWrong is true.
2. She is almost, but not completely, sure that eating meat is not wrong in her exact circumstances.
3. She is sure that eating vegetables is not wrong in her exact circumstances.
4. She is sure that she has states 1–3.

A little reflection shows that this is an incoherent set of states. Given ProbWrong, it is simply wrong for someone with states 2 and 3 to eat meat. And the agent knows that she has states 2 and 3. So she can deduce from her other commitments and mental states that eating meat is, right now, wrong. So she shouldn't be almost sure that eating meat is not wrong; she should be sure that it is wrong.

This argument generalises. If 1, 3 and 4 are true of any agent, the only ways to maintain coherence are to be completely certain that meat eating is not wrong, or completely certain that it is wrong. But that is absurd; these are hard questions, and it is perfectly reasonable to be uncertain about them. At least, there is nothing incoherent about being uncertain about them. But ProbWrong implies that this kind of uncertainty is incoherent, at least for believers in the truth of ProbWrong itself. Indeed, it implies that in any asymmetric moral risk case, an agent who knows the truth of ProbWrong and is aware of her own mental states

cannot have any attitude between certainty that both options are not wrong, and certainty that the risky action is not, in her exact circumstances, wrong. That is absurd.

I conclude that any version of the normative internalist thesis where N1 = N2 is also absurd. Happily, that view seems to be shared by existing defenders of internalism, who usually defend versions where N1 $\neq$ N2. So I'll set the N1 = N2 versions of internalism aside and focus just on the versions where they come apart.

2.5. Motivation One: Guidance

The externalist offers a fairly simple piece of advice to people facing a moral challenge: Do the right thing. But as a general piece of advice, *Do the right thing* might sound not much more helpful than *Buy low, sell high*. We need, it might be thought, more helpful advice.

That kind of consideration plays a big role in our thinking about factual uncertainty. Think again about Botum the charity director. The best outcome for her, and for the cause she is working for, would be for her to bet the £10,000 on the number that will actually win. But we don't think she's under an obligation to do that. Indeed, we think she is under an obligation to not even try to do that. One reason for that, arguably, is that the strategy *Bet on the winning number* is not one she is in a position to carry out.

Now the externalist does think that agents should carry out the strategy *Do the right thing*. But in cases where the moral evidence is murky, arguably this is no more a reasonable demand than the demand that Botum bet on the winning number. Here is how Michael Smith puts the point. He has just rehearsed Frank Jackson's argument, involving cases like Billie, for the conclusion that right action does not involve maximising the probability of the best outcome, but maximising expected value.

> Indeed, anyone impressed by Jackson's argument on the non-evaluative facts side of things should surely suppose that an equally impressive argument could be made for the conclusion that right action consists not in the maximization of expected *value*, but rather in the the maximization of expected *value-as-the-agent-sees-things*. For no mere exercise of such capacities as an agent has looks like it will ensure that what is really valuable will manifest

> itself to her either. There are, after all, cultural circumstances in which it would be wildly optimistic to suppose that agents could, merely through the exercise of their own rational capacities, come to judge to be valuable what's really valuable ... If this is right, however, then it seems that the most that we could ever expect of a normal agent ... is that they form their evaluative commitments in a way that is sensitive to such evidence as is available to them and that they form their desires in a way that is sensitive to their evaluative commitments. (M. Smith 2006, 143)

Andrew Sepielli expresses a similar sentiment.

> The problem is that we cannot base our actions on the correct normative standards; our relationship to such standards is limited to mere conformity to them. This follows from a quite general point—that we cannot guide ourselves by the way the world is, but only by our representations of the world. (Sepielli 2009, 8)

And we saw in the previous chapter that similar sentiments are expressed by Ted Lockhart (2000, 8–9), William MacAskill (2014, 7) and by Hillary Greaves and Toby Ord (2017). We might try to turn this idea into an argument for internalism as follows.

1. Our most important norms should be sources of usable advice.
2. If normative externalism is true, our norms are not sources of usable advice.
3. If normative internalism is true, our norms are sources of usable advice.
4. So normative externalism is false, and we have a reason to believe normative internalism is true.

Note that I'm not here assuming that normative externalism and normative internalism are contradictories; there are positions that might best be classified as falling into neither camp. If they were contradictories, the second conjunction of the conclusion would be highly redundant.

One problem for this argument is that it relies on a slippery notion of usability. If we have rather generous standards for what counts as a usable norm, then premise 2 of the argument is false. After all, we can often tell what is the right thing to do. If we have rather strict standards, then premise 1 is false, since it amounts to the claim that the application conditions for the most important

norms must be luminous. (A norm is luminous if whenever it applies, it is possible to know that it applies.) But Timothy Williamson (2000) has shown that nothing interesting is luminous, and our most important norms are interesting. I suspect that there is no reading of 'usable' that makes both premises 1 and 2 true.

The slipperiness also extends to premise 3. The internalist needs standards that are usable, in their preferred sense, and which Robespierre violates. (Unless they are happy saying that Robespierre did well, in the sense that's most important to them.) But they need that sense of usability to be one in which *Do the right thing* is not usable. And it is hard to see what that sense could be.

The regress arguments that will recur throughout this book are designed, in part, to back up this conclusion. (See particularly the discussion of inter-theoretic value comparisons in section 6.2.) I'm going to be arguing that everyone except the most radical subjectivist will be have to acknowledge standards for evaluating agents that those very agents are not in a position to accept. The only options, I'll argue, are radical subjectivism, and norms that are not guaranteed to be able to usable in the internalist's preferred sense. That is, the norms will only be usable in the sense that *Do the right thing* is usable. Since this radical subjectivism is false, some monsters really do well by their own lights, the connection between evaluation and guidance must be more tenuous than the internalist assumes.

2.6. Motivation Two: Recklessness

A different argument against externalism is that it licences a form of moral recklessness. And this kind of moral recklessness should not be licenced, says the objector, it should be condemned.

To see the problem, start with the example of Deòrsa, the uncertain carnivore. (This case is discussed by Guerrero (2007), who uses it in mounting an attack on moral recklessness.) And let's assume that Deòrsa does end up deciding that he will eat meat. Deòrsa knows that the moral risks are largely, if not universally, on one side. He knows that eating meat provides him with just a small benefit, but puts him at risk of being a moral monster. And yet he does it.

Now by hypothesis, a fully informed agent in Deòrsa's position would do the same thing. And yet it is easy to feel some unease with the externalist verdict

that Deòrsa's actions are right, rational and blameless. There is a whiff of recklessness about Deòrsa's actions, and this kind of recklessness may seem to be a moral vice.

We can make this whiff stronger by tightening the analogy with reckless action. For example, imagine that Deòrsa isn't mostly certain that meat eating is acceptable. In fact, in the revised case he is very confident that meat eating is wrong. And yet, he eats meat anyway. The analogies between Deòrsa and Cressida, the reckless driver, start to feel compelling at this point. And yet the externalist says that Deòrsa is not doing anything wrong, or irrational, or blameworthy. (This variant, and its importance, was suggested by Andy Egan.)

Or perhaps we can build the analogy directly into Deòrsa's case. Imagine that as well as choosing what to eat, Deòrsa is choosing how to cook it. Deòrsa is considering trying out a new technique from a modernist cookbook. He knows that a side effect of this technique is that a distinctive kind of chemical is released into his building's ventilation. This chemical will build up in large quantities in his apartment and the apartment next door. The chemical is odorless, and harmless to everyone who doesn't have a particular allergy. But the quantities Deòrsa would release would be fatal to anyone with the allergy. And Deòrsa knows the boy in the next apartment has some kind of rare allergy, though he can never remember which one it is. He thinks it is probably some other allergy the boy has, and in fact he is right. So he cooks the meat using the modernist technique.

To make the analogy explicit, assume that Deòrsa has equal credence in these two propositions.

1. Meat eating is morally acceptable.
2. The boy in the next apartment will not have a fatal reaction to the chemical that will be released by the modernist cooking technique.

In each case, this credence is high, but far from maximal. Unless Deòrsa knows that 2 is true, what he does is horribly reckless. It's not worth risking killing one of your neighbours to get the benefits of a new method of meat preparation. Similarly, says the internalist, the gustatory benefits of meat aren't worth the risk that goes along with joining the meat-eating team.

D. Moller (2011) similarly argues that internalism is motivated by considerations about recklessness. I'll respond to Moller's own example at more length below,

so let me start with my own variant of the kind of case that motivates his position. Two CEOs are trying to choose between more aggressive and more conservative business strategies. Each commissions internal inquiries to determine some properties of the aggressive strategy. (They know how conservative strategies work, since those strategies are familiar.) One of the CEOs doesn't know exactly what the practical consequences of the aggressive strategy will be, so she commisions an inquiry into those practical consequences. And the other CEO doesn't know what the right moral evaluation of the aggressive strategy is, so he commissions an inquiry into its moral evaluation. Both inquiries come back with a 3–2 split. In the first case, all five agree the aggressive strategy will slightly raise profits relative to the conservative strategy. But two members think that a side effect will be that ten people in nearby communities fall sick and die as a consequence of the company's operations. In the second case, all five think the conservative strategy is morally acceptable. But three think the aggressive strategy is good enough, while the other two think it is as bad as being responsible for ten avoidable deaths. In each case, it turns out, the majority members of the committee are right, though the CEO has no extra evidence for that. The intuition these cases seem to support is that neither CEO should carry out the aggressive strategy. Indeed one might hold (though Moller, interestingly, does not) that we should think of the CEO's who carry out these strategies as being equally culpable for their recklessness.

2.7. Motivation Three: Symmetry

Both the guidance considerations and the recklessness considerations push one towards thinking that factual uncertainty and moral uncertainty should be treated symmetrically, or at least as symmetrically as possible. I briefly mentioned that Moller expressly rejects the symmetry claim, and the failure of N1=N2 versions of internalism make it hard, at least for non-consequentialists, to endorse perfect symmetry. But there is something to the idea that moral uncertainty and factual uncertainty should get very similar theoretical treatments, and the externalist offers very different theoretical treatment of them.

We could get to this idea in a few ways. We could try to argue that it follows from considerations about guidance or recklessness. We could try to argue that it best explains intuitions about guidance or recklessness. Or we could just argue for it directly, either my appeal to the intuitive plausibility of the symmetry claim, or the intuitive plausibility of what it says about a number of cases. For instance,

we could just argue that it is plausible that whatever negative attitude we have towards Cressida's actions, and to Cressida, we should have towards Deòrsa's actions, and to Deòrsa. And we could argue that whatever positive attitude we have towards Billie's actions, and to Billie, we should have to the person who successfully manages to maximise evidential expected goodness. In short, we should have symmetric attitudes about the philosophical significance of normative uncertainty and factual uncertainty.

Ths idea that symmetry (or near-symmetry) should be built into our theories will not, I suspect, strike most people as absurd. Indeed, I suspect it strikes many people as so plausible it barely needs defence. It certainly does a lot of work, without much argument, in works by Jacob Ross (2006) and Michael Zimmerman (2008). If the symmetry thesis is both intuitive and true, there's nothing wrong with this approach. And I concede it is, at least prima facie, highly intuitive. But I don't think it is true. Indeed, I don't think it is even particularly intuitive, once we reflect on it in more detail.

But it is intuitive enough to use as the foundation for discussions of internalism. And while I'll cycle back around to other motivations for internalism, I'll use symmetry-based considerations as the main focus of discussion. That's because the symmetry-based considerations do such a good job of both being independently intuitive, and capturing what is best worth capturing in the other arguments.

So in the next chapter I'll push back against the intuitiveness of this symmetry claim, arguing that the closer we look at it, the less similar factual and moral uncertainty seem. And in the chapter after that, I'll argue that even if symmetry is plausible it should be rejected, for it leads to unacceptable regresses.

3. Against Symmetry

In the previous chapter, I suggested that one of the key motivations for normative internalism that it allows for a symmetry between the way we treat factual uncertainty and ignorance, and the way we might think about treating normative uncertainty and ignorance. Some writers have found it so obvious that these cases should be treated symmetrically that they have simply incorporated this symmetric treatment into their theory without arguing for it. Those who have argued for it have usually found the symmetry very intuitive.

In this chapter, I'll try to undermine that intuitive symmetry. The first three sections will introduce three considerations that undermine the idea that the factual and normative uncertainty should be treated symmetrically, and the last three sections deal with some complications that the first three sections introduce. In the next chapter, I'll argue that even if we found the symmetry intuitive, we should ultimately reject it, because there is no way to incorporate it into a theory that is even remotely plausible. That is, I'll argue that any internalist theory that can handle even very simple cases has to reject the symmetry thesis, and so cannot be motivated by symmetry considerations.

3.1. Guilt and Shame

If normative and factual uncertainty have the same normative implications, then we should feel similarly about our own past actions that were done due to factual ignorance, and those that were done due to moral ignorance. But this doesn't seem to be how we do, or should, feel. We can see this by comparing a pair of cases. The second of the cases is a minor modification of a case Elizabeth E. Harman (2015) uses in making a similar argument to the one I'm presenting in this section.

Prasad is a father of two children, an older daughter and a younger son. In the division of parental labour in his house, teaching the children to read is primarily his responsibility. He takes this very seriously, and reads the latest studies

on which techniques are most effective at teaching reading. He doesn't have a strong enough background in statistics to be able to evaluate many of the papers he reads, but he can tell what techniques are being approved by the leading figures in the field, and those are the techniques he uses in teaching his children to read.

Unfortunately, the relevant science around here moves slowly and fitfully. The technique that Prasad followed when his daughter was learning to read was soon shown to be mostly ineffective. It was better than not spending time on reading, but wasn't any better than unstructured reading time. By the time his son was learning to read, educational science had advanced substantially, and Prasad was able to use a technique that led to his son learning to read relatively quickly. This gave his son an advantage that persisted throughout his schooling, and led to him being admitted to an exclusive college, and subsequently earning much more than he would have without the benefit of early reading. Prasad's daughter did well at school, as you'd expect with this level of parental attention, but would have been even better off had been trained to read the way her brother was trained.

Archie is a 1950s father who, like many other 1950s fathers, thinks it is more important to look after his son's interests than his daughter's. So while he puts aside a substantial college fund for his son, he puts aside less for his daughter. As a consequence, his daughter cannot afford to go to as good a college as his son goes to, and subsequently is materially less well off throughout her life than Archie's son.

Prasad was mistaken about a matter of fact; about which techniques are most effective at teaching a child to read. Archie was mistaken about a moral matter; whether one should treat one's sons and daughters equally. Now consider what happens when both see the error of their ways. Prasad may feel bad for his son, but there is no need for any kind of self-reproach. It's hard to imagine he would feel ashamed for what he did. And there's no obligation for him to feel guilty, though it's easier to imagine him feeling guilty than feeling ashamed. Archie, on the other hand, should feel both ashamed and guilty. And it's natural that a father who realised too late that he had been guilty of this kind of sexism would in fact feel the shame and guilt he should feel. The fact that his earlier sexist attitudes were widely shared, and firmly and sincerely held, simply seems irrelevant here.

If the symmetry thesis were correct, there should not be any difference in Prasad and Archie's attitudes. Both of them behaved in just the way we should expect,

given their factual and normative beliefs. And both of them had beliefs that were sincere, and widely shared in their community. But there is still a difference between the two of them, as revealed by the emotional reactions they both do and should have.

3.2. Jackson Cases

As Zimmerman (2008) argues, the kinds of cases discussed by Jackson (1991) are important for seeing how factual uncertainty is normatively significant. It isn't just that when an agent doesn't know what is true, and so doesn't know which action produces the best outcome, she thereby doesn't know what is right to do. In some cases of decision making under uncertainty, the thing that is clearly right to do is the one thing she knows will not produce the best outcome. Gambling the charitable donation on the roulette wheel is wrong, although the best outcome would be to gamble on the number that will actually come up. In the previous chapter we dubbed cases like this, where the right thing to do is something one knows will not produce the best outcome, Jackson cases. Jackson cases are ubiquitous when making decisions under factual uncertainty.

If we should treat factual uncertainty and moral uncertainty symmetrically, then Jackson cases for moral uncertainty would be easy to find. But it is far from clear that there are any such cases. That is, it is far from clear that there are cases where we want to say anything positive about an agent who hedges their moral bets.

A simple way to generate Jackson cases is to set up a decision problem with the following features:

- There are three option: A, B and C;
- There are two epistemic possibilities, w_1 and w_2, the agent knows that precisely one of them is realised, and the she reasonably thinks each is fairly likely.
- In w_1, A is optimal, C is a little worse, and B is a catastrophe.
- In w_2, B is optimal, C is a little worse, and A is a catastrophe.

If the agent's uncertainty about w_1 or w_2 is grounded in a straightforwardly factual uncertainty, it seems the agent should do C. Just what that 'should' amounts to is up for debate, but there is something awful about doing A or B - even if it produces the optimal outcome.

What happens, though, if w_1 and w_2 are factually alike, but differ in the correct moral theory? (As has come up a few times, it is unlikely that both w_1 and w_2 will be *possible* worlds in this case, but I don't think this matters for current purposes.) Well, let's look at some cases and see.

3.2.1. Case One - Abortion

Marilou is 12 weeks pregnant, and lives in a state where abortion is criminalised and, on occasion, heavily punished. Marilou deeply desires to have an abortion. Marilou is reasonably well off, and as is the norm in states that criminalise abortion, reasonably well off people are able to obtain abortions with a little assistance. Marilou asks her friend Shila for such assistance. Shila now has to make a choice. Shila is torn between two moral views about abortions 12 weeks into pregnancy. According to one, the potential that the fetus has to develop into a fully functioning human being means that aborting it is the moral equivalent of murder. According to another, the fetus has little or no moral standing on its own, the importance of Marilou's autonomy means that Marilou should be able to get an abortion, and her friends should assist her in avoiding the oppressive laws against abortion. Shila now has three choices.

1. Assist Marilou in getting the abortion, which is either a way of respecting Marilou's autonomy and honouring their friendship, or is a way of being an accomplice to murder.
2. Report Marilou's plans to the authorities, which is either horribly disrespectful to Marilou and a gross violation of their friendship, or bravely preventing a murder. (Assume that Shila knows that although the authorities aren't maximally vigilant about preventing abortions, they are obliged to act on incriminating information, so this tip-off will lead to Marilou's imprisonment.)
3. Do nothing, suspecting that without her help, Marilou will carry the child to term and quietly adopt it out.

In either w_1, the world where abortion is permissible, or w_2, the world where it is not, C is bad. In w_1, Shila is a bad friend, and is tacitly collaborating in state oppression. In w_2, Shila is not taking simple steps that would remove the mortal danger facing an innocent human. But option C isn't catastrophic in either world. In w_1, Shila is not personally stopping Marilou get an abortion, she just isn't helping Marilou break the law. (You can be a good enough friend and still draw the line between helping one move houses and helping one move

bodies.) And in w_2, she's not killing anyone, or even letting someone be killed, just not being maximally vigilent in preventing a killing. So the case has the structure of a Jackson case.

And yet there is little to be said for C. The situation calls for moral bravery, one way or the other. (I think in the direction of A, but it doesn't matter for these purposes whether you agree with that.) And C is moral cowardice. Unlike in the cases involving factual uncertainty, it doesn't seem at all like the safe, prudent, commendable option.

3.2.2. Case Two - Theft

Eurydice and Pandora are acquaintances, and they are planning to go to a party. Eurydice is worried because Pandora plans to wear some very expensive jewellery, and the party features a number of thieves, several of whom are Eurydice's friends. Eurydice tells Pandora this, but Pandora is unmoved, and insists she won't be deterred from living her life the way she wants by the existence of petty criminals. Eurydice is much more observant than Pandora, and knows that if someone tries to steal the jewellery, she'll be able to prevent them, but only by using a non-trivial amount of physical force. For example, she could punch the would-be thief hard in the jaw while he was making his escape, revealing his thievery. (Realistically, she can't know exactly how she would prevent a theft, but assume that's the level of force that would be needed.)

Eurydice is torn between two moral theories. One of them is a fairly mainstream view on which a moderate amount of physical force is warranted if it is the only way to prevent the theft of expensive goods. On the other moral theory, the demands of friendship and bodily autonomy completely outweigh considerations arising from property, so punching a friendly thief to prevent a theft would be a completely unjustified assault. Given all this, Eurydice has three options.

1. Go to the party and plan to prevent (using violence if necessary) any theft of Pandora's jewellery.
2. Go to the party and plan to refrain from any violence, even if this means standing by while a theft occurs.
3. Prevent Pandora going to the party. The most morally acceptable way to do that, Eurydice thinks, would be to tell Pandora a small lie that leads to Pandora going on a wild goose chase for half the night, leaving it impossible to go to the party.

Again, this feels like a Jackson case. C is a moral misdemeanour - you shouldn't lie to people for the purpose of distracting them away from a party they have every right to be at. But it's worse to stand by and watch a theft take place that you could easily (and properly) prevent, or to unjustifiedly punch a friend in the jaw.

Yet again it seems like C would be a terrible option to take. Either the amount of violence needed to apprehend the thief would be justified or it wouldn't be. In neither case does it seem like sending Pandora on a wild goose chase to prevent the theft would be a good way to prevent the problem arising. This seems true even though it would guarantee that things don't go badly morally wrong, while either alternative runs a substantial moral risk.

3.2.3. An Asymmetry

When welfare is on the line, it is not just acceptable, but laudable, to sacrifice the chance of the best outcome for a certainty of a very good outcome. But it isn't at all clear that this is true when virtue is on the line. Committing a moral misdemeanour because you don't know which of the other options is a moral felony and which is the right thing to do is, still, committing a moral misdemeanour.

3.3. Motivation

Moral uncertainty, at least of the kind I'm focussing on, is a kind of constitutive uncertainty. An agent who is morally uncertain is uncertain about what kind of things constitute goodness, rightness, praiseworthiness, and so on. It's very plausible that these are indeed constituted by something else. It's hard to imagine that rightness is a free-floating feature of reality.

Cases of constitutive uncertainty are useful test cases for thinking about what's really valuable. If we know that A constitutes B, and hence have equally strong desires for A and for B, it isn't always easy to tell which of these desires is more fundamental, and which is derived. Of course, neither of the desires will be an *instrumental* desire, since getting A isn't a means to getting B. But one of them could be derivative on the other.

And the simplest way to tell which is which, is to look to people who do not know that A constitutes B, and see what makes sense from their perspective. Think again about Monserrat, who has forgotten the victory conditions for her game.

We know that being first to 10 points constitutes winning. But she doesn't. What action makes sense for her to do? I think it is doing the thing that maximises her probability of winning, given her credal distribution. It turns out that isn't the thing that maximises her probability of being first to 10 points, which is what actually amounts to winning. But she has no motivation to be first to 10 points, unless that amounts to winning. Or, at least, she has no such motivation on the most natural telling of the story. Perhaps she has an odd psychological tick that means she always values being first to *n* figures in points in any game she plays. But the more natural story is that she wants to win, and she should do the thing that maximises the probability of winning.

Things are rather different when it comes to moral uncertainty. There it seems that agents should be moved to produce the outcome that actually constitutes goodness or rightness, not the thing that maximises expected goodness or rightness. This is a point well made by Michael Smith. He compared the person who desires to do what is actually right, as he put it, desires the right de re, with the person who desires to do what is right whatever that turns out to be, as he put it, desires the right de dicto.

> Good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not just one thing: doing what they believe to be right, where this is read *de dicto* and not *de re*. Indeed, commonsense tells us that being so motivated is a fetish or moral vice, not the one and only moral virtue. (M. Smith 1994, 75)

I think that's all true. A good person will dive into a river to rescue a drowning child. (Assuming that is that it is safe enough to do so; it's wrong to create more rescue work for onlookers.) And she won't do so because it's the right thing to do. She'll do it because there's a child who needs to be rescued, and that child is valuable.

Not everyone agrees with Smith that commonsense has this verdict about moral motivation. It helps to see the point made less abstractly, about a particular case. Here is the initial description of Saint-Just from Palmer's classic study of the Committee of Public Safety, *Twelve Who Ruled*.

> Saint-Just was an idea energised by a passion. All that was abstract, absolute and ideological in the Revolution was embodied in his slender figure and written upon his youthful face, and was made

> terrible by the unceasing drive of his almost demonic energy. He was a Rousseauist, but what he shared with Rousseau was the Spartan rigor of the *Social Contract*, not the soft day-dreaming of the *Nouvelle Héloïse*, still less the self-pity of the *Confessions*. He was no lover of blood, as Collot d'Herbois seems to have become. *Blood to him simply did not matter*. The individual was irrelevant to his picture of the world. The hot temperament that had disturbed his adolescence now blazed beneath the calm exterior of the political fanatic. (Palmer 1941, 74, emphasis added)

That's what someone who is only motivated by the good, as such, looks like. And it's terrifying. Commonsense morality prefers a view where blood matters, and the individual is relevant, and where all of Rousseau's works have something to teach us about how to live. [1]

We need to distinguish here two theses one might have about moral motivation. One is that the good, as such, should not be one's only motivation. That's what Smith says commonsense says, and it's what the example of Saint-Just supports. Another is that the good, as such, should not be among one's motivations. I think this latter claim is mostly true as well. But I'll come back to that; for now I want to spell out the consequences of the weaker claim, that the good should not be one's only motivation.

This claim already makes trouble for normative internalists, including Smith himself. It makes trouble because it offers us a nice explanation of why there should be the kind of asymmetry between factual and normative uncertainty that we see in cases like Shila's. Think again about the situation she is facing. She has to choose between respecting Marilou's autonomy, and respecting the foetus's life. And she doesn't know what to do, in no small part because she doesn't know which form of respect constitutes moral rightness. But one thing she does know is that the moderate option maximises expected goodness. If we thought that this was an important motivation, we presumably should think it could be decisive in some cases, and Shila might take that moderate option. But intuitively she should never do that, and not have any motivation to do that. A

[1] I've mentioned Robespierre a few times in this context, so it's interesting to note that Palmer thinks Robespierre is not as extreme as Saint-Just. He compares the two in the paragraph preceding this one, mostly saying that Saint-Just is a more extreme version of Robespierre. Saint-Just is similar to his hero, but "without the saving elements of kindness and sincerity". I think 'saving' is a little strong, but otherwise that judgment seems right. Collot positively desired actually bad things, Robespierre cared insufficiently about actually good things, and Saint-Just simply did not care about anything beyond ideology.

pro-choice theorist may think Shila should believe that respecting autonomy is the right thing to do, and so Shila should be motivated to do what's right because she's motivated to respect autonomy. A pro-life theorist may think Shila should believe that respecting life is the right thing to do, and so Shila should be motivated to do what's right because she's motivated to respect life. But neither will hold that Shila should have a motivation to do what's right that floats free of her motivation to respect autonomy, and to respect life.

This way of thinking about Shila's case suggests a prediction, one that is borne out by the cases. It isn't always the case that moral 'hedging', of the kind I've been criticising since the start of section 3.2, is bad. Imagine an agent faces a choice between competing values, both of which are values that she holds dear. For instance, consider an administrator who faces a student in a somewhat unusual situation. (The point of it being unusual is to ensure there is no clear precedent for what to do in such cases.) The administrator has to choose between being compassionate to the person in front of her, and doing the thing she thinks would best treat the case in front of her like previous cases. She may well care both about compassion and equality, and in such a case, it would make sense to look for a way to minimise the distance between how she treats this case and how she has treated past cases, while also being highly compassionate to the person in front of her. And that is true even if the outcome she comes up with is neither the most compassionate thing she can do, nor the most respecting of her desire to treat like cases alike. The reason this makes sense is that the administrator doesn't think rightness is either exclusively constituted by compassion, or by treating like cases as alike as possible. Rather, she has plural values, like most of us do. And plural values, as opposed to uncertainty about what is the one true value, can produce moral Jackson cases.

What is the difference between Monserrat's case and Shila's? Why should Monserrat aim for what maximises the constituted quantity, while Shila aims for what maximises (or perhaps best respects) the constituting quantity? The answer comes from what it means for something to be right. It just is for it to be valuable. One of the striking things about games is that they turn something otherwise pointless, like being first to 10 points, into something that rational people can value. But morality isn't like that. It can't make value out of something that wasn't valuable, because if it wasn't valuable, it wouldn't be fit to constitute rightness. So whatever rightness is, be it respecting autonomy or maximising welfare or whatever, must be something already valuable. And it is hard to see how having the property of being most valuable can be more valuable than the valuable thing itself.

So we get an explanation of Smith's observation. (And here I'm not saying anything that hasn't been said before, by for example Nomy Arpaly (2003) and Julia Markovits (2010).) It is good to aim at what is actually right and good, not at rightness and goodness themselves, because the constitutors are where the value lies. But that means moral uncertainty should not affect our motivations. And that's a striking asymmetry with factual uncertainty, which quite clearly should affect our motivations.

3.4. Welfare and Motivation

Smith's insight, that there is something wrong about being motivated to do what's good as such, generalises. There are plenty of other things where we do and should care about their constituents, but we should not (and typically do not) care about them as such. Welfare, for instance, is like this.

It's plausible that deliberately undermining your own welfare, for no gain of any kind to anyone, is irrational. Indeed, it may be the paradigmatic form of irrationality. There is a radically Humean view that says that welfare just consists of preference satisfaction, and rationality is just a matter of means-end reasoning. If that's right then what I just said is plausible is not only true, but almost definitional of rationality. You don't have to be that radical a Humean, or really any kind of Humean at all, to think there is a connection between welfare and rationality. But if rationality is connected to welfare, it is because it is connected to the constituents of welfare, not to welfare as such. To see this, consider two examples, Bruce and Oberon.

Bruce has thought a bit about philosophical views on welfare. In particular, he has spent a lot of time arguing with a colleague who has the G. E. Moore-inspired view that all that matters to welfare is the appreciation of beauty, and personal love.[2] Bruce is pretty sure this isn't right, but he isn't certain, since he has a lot of respect for both his colleague and for Moore.

Bruce also doesn't care much for visual arts. He thinks that art is something he should learn something about, both because of the value other people get from art, and because of what you can learn about the human condition from it. And while he's grateful for what he learned while trying to inculcate an appreciation

[2] It would be a bit of a stretch to say this is Moore's own view, but you can see how a philosopher might get from Moore (1903) to here. Appreciation of beauty is one of the constituents of welfare in the objective list theory of welfare put forward by John Finnis (2011, 87–88).

of art, and he has become a much more reliable judge of what's beautiful and what isn't, the art itself just leaves him cold. I suspect most of us are like Bruce about some fields of art; there are genres that we feel have at best a kind of sterile beauty. That's how Bruce feels about visual art in general. This is unfortunate; we should feel sorry for Bruce that he doesn't get as much pleasure from great art as we do. But it doesn't make Bruce irrational, just unlucky.

Finally, we will suppose, Bruce is right to reject his colleague's Moorean view on welfare. Appreciation of beauty isn't a constituent of welfare. We'll for the sake of the example that welfare is a matter of health, happiness and friendship. That is, a fairly restricted version of an objective list theory of welfare is correct in Bruce's world. And for people who like art, appreciating art can produce a lot of goods. Some of these are direct - art can make you happy. And some are indirect - art can teach you things and that learning can contribute to your welfare down the line. But if the art doesn't make you happy, as it doesn't make Bruce happy, and one has learned all one can from a genre, as Bruce has, there is no welfare gain from going to see art. It doesn't in itself make you better off, in the way that Bruce's Moorean colleague thinks it does.

Now Bruce has to decide whether to spend some time at an art gallery on his way home. He knows the art there will be beautiful, and he knows it will leave him cold. There isn't any cost to going, but there isn't anything else he'll gain by going either. Still, Bruce decides it isn't worth the trouble, and stays out. He doesn't have anything else to do, so he simply takes a slightly more direct walk home, which (as he knows) makes at best a trifling gain to his welfare.

Bruce is perfectly rational to do this. He doesn't stand to gain anything at all from going to the gallery. In fact, it would be a little perverse, in a sense we'll return to, if he did go.

Oberon is also almost, but not completely certain, that health, happiness and friendship are the sole constituents of welfare.[3] But he worries that this is undervaluing art. He isn't so worried by the Moorean considerations of Bruce's colleagues. But he fears there is something to the Millian distinction between higher and lower pleasures, and thinks that perhaps higher pleasures contribute more to welfare than lower pleasures. Now most of Oberon's credence goes to alternative views. He is mostly confident that people think higher pleasures are more valuable than lower pleasures because they are confusing causation and constitution. It's true that experiencing higher pleasures will, typically, be

[3]Thanks to Julia Markovits for suggesting the central idea behind the Oberon example, and to Jill North for some comments that showed the need for it.

part of experiences with more downstream benefits than experiences of lower pleasures. But that's the only difference between the two that's prudentially relevant. (Oberon also suspects the Millian view goes along with a pernicious conservatism that values the pop culture of the past over the pop culture of the present solely because it is past. But that's not central to his theory of welfare.) And like Bruce, we'll assume Oberon is right about the theory of welfare in the world of the example.

Now Oberon can also go to the art gallery. And, unlike Bruce, he will like doing so. But going to it will mean he has to miss a night playing video games that he often goes to. Oberon knows he will enjoy the video games more. And since playing video games with friends helps strengthen friendships, he has a further reason to skip the gallery and play games. Like Bruce, Oberon knows that there can be very good consequences of seeing great art. But also like Bruce, Oberon knows that none of that relevant here. Given Oberon's background knowledge, he will have fun at the exhibition, but won't learn anything significant.

Still, Oberon worries that he should take a slightly smaller amount of higher pleasure rather than a slightly larger amount of lower pleasure. And he's worried about this even though he doesn't give a lot of credence to the whole theory of higher and lower pleasures. But he doesn't go to the gallery. He simply decides to act on the basis of his preferred theory of welfare, and since that theory of welfare is correct, he maximises his welfare by doing this.

Now distinguish the following two claims about welfare and rationality. The first of these claims is plausibly true; the second is false.

- A person's welfare is such that it is irrational for them to do something that might undermine it for no compensating gain.
- It is irrational for a person to do something that might undermine their welfare, whatever that turns out to be, for no compensating gain.

If welfare turns out to be health, happiness and learning, then the first claim says that it is irrational to risk undermining one's health, happiness and learning for no compensating gain. And that is correct. But the second claim says that for any thing, if that thing might be welfare, and an action might undermine it, it is irrational to perform the action without a compensating gain. That's a much stronger, and a much less plausible, claim. The examples of Bruce and of Oberon show that it is false; they act rationally even though they do things that might undermine what welfare turns out to be.

One caveat to all this. On some theories of welfare, it will not be obvious that even the first claim is right. Consider a view (standard among economists) that welfare is preference satisfaction. Now you might think that even the first claim is ambiguous, between a claim that one's preferences are such that it is irrational to undermine them (plausibly true), and a claim that it is irrational to undermine one's preference satisfaction. The latter claim is not true. If someone offers a person a pill that will make her have preferences for things that are sure to come out true (she wants the USA to stay being more populous than Monaco, she wants to have fewer than ten limbs; etc.), it is rational to refuse it. And that's true even though taking the pill will ensure that she has a lot of satisfied preferences. What matters is that taking the pill does not satisfy her actual preferences. If she prefers X to Y, she should aim to bring about X. But she shouldn't aim to bring about a state of having satisfied preferences; that could lead to rather perverse behaviour, like taking this pill.

3.5. Motivation, Virtues and Vices

So far in this chapter I have relied heavily on Michael Smith's principle that a certain kind of motivation would be unreasonably fetishistic. In this section I'm going to defend Smith's principle in more detail. Since Smith's principle has been extensively discussed, I'm going to spend some time on the existing literature. But one key point of this section will be that I need a much weaker principle for my broader conclusion than Smith needs for his. So even if the existing objections to Smith are correct, and I will concede at least one has some force against the strong principle Smith defends, they may not affect my argument for externalism.

That Smith and I need different versions of the principle should not be too surprising. As we saw in chapter 2, Smith defends some of the internalist principles I'm arguing against. Since we have different conclusions, one might hope we had different premises. The passage from Smith I quoted about moral fetishism is in defence of his motivational internalism. As I noted in chapter 1, the different theses called internalism are dissociable, but they do have some affinities. Motivational internalism is consistent with normative externalism, but is in some tension with it. So again, it isn't surprising that I'll be using Smith's idea in a slightly different way.

Let's start by setting out three theses that one might try to draw from considerations starting from Smith's reflections.

Weak Motivation Principle (WMP) In equilibrium, it is permissible to not be intrinsically motivated by maximally thin moral properties de dicto.

Strong Motivation Principle (SMP) In most circumstances, it is impermissible to be at all intrinsically motivated by moderately thin (or thinner) moral properties de dicto.

Ideal Motivation Principle (IMP) In all circumstances, it is impermissible to be at all intrinsically motivated by maximally thin moral properties de dicto.

The SMP and IMP are both stronger than the WMP, though neither is stronger than the other. As I read him, Smith needs the IMP to get his argument for motivational internalism to work. Since I'm not interested in that, I'll set it aside from now on.

In the next section I'll discuss the WMP, with a focus on clarifying the term 'equilibrium'. The aim is to argue that it is is true, and that if it is true, there is an asymmetry between factual and moral uncertainty.

After that, I'll discuss the SMP. I also think the SMP is true, and if it is true, then there is a huge asymmetry between factual and moral uncertainty. But I need to stress at this point that defending the SMP isn't strictly necessary for the major argument of the chapter; the WMP is enough to raise problems.

After that, I'll discuss a few examples that help clarify the boundaries of the two principles, and which I think provide some argument for the principles. But I'm discussing them at the end, because I don't really want the case for or against the principles to rest on intuitions about disputed examples like the ones I'll bring up.

The principles appeal to the notion of 'intrinsic motivation', and it's worth spending a few words on that. Just about everything I say here is drawn from Arpaly and Schroeder (2014, 6–14), and they go into more detail than I do about some of the important distinctions.

There is a distinction in everyday English between ends and means. And to a first approximation, to desire something as an end is to desire it intrinsically, and to desire it as a means it to desire it instrumentally. But here we need to make a slightly finer distinction than that.

Parents typically desire that their children be well educated. For some people this will be an instrumental desire; they want their children to be, say, very rich, and think that education is a means to wealth. But for others it will be intrinsic; a good education is part of what is good for their children.

Now consider the desire (again widely held among parents) that one's children be well educated in arithmetic. How does this relate to the general desire that they be well educated? It isn't exactly a means to that end. It is part of what it is to be well educated. To desire that a child be well educated, and to know what it is to be well educated, just means that you desire that the child be well educated in arithmetic. Call desires like this, ones which have a constitutive rather than causal connection to intrinsic desires, *realizer* desires.

The most obvious cases of realizer desires are when the intrinsic desire is more general, and the realizer desire is more specific. But we can go the other way around too. Consider again the perfectly normal parent who wants their child to be well educated, to be healthy, to be happy, to have lots of friendships, and generally wants all the things that make up a good life for their child. That parent will want their child to have a good life. This might be an intrinsic desire; maybe all those other desires are realizers of it. It might even be an instrumental desire, though this would be a little perverse. Or it might be a realizer desire, and I think this is the most natural case. If one wants the child to be happy, healthy, befriended, educated, etc, and one has a sensible balance between those desires, then in virtue of all that, one has the desire that the child have a good life. To desire all these things just is to desire the child have a good life. It's a very different way of desiring that the child have a good life than having that desire instrumentally, as one might if one wanted the child to have a good life solely so one would be rewarded in the afterlife. And it is a somewhat different way of desiring that the child have a good life than having that desire intrinsically. The difference shows up in two ways. One concerns the order of explanation: does one want the child to have a good life in virtue of wanting the child to be happy, healthy etc, or is it the other way around? The other concerns how one's desires for the child change when one's conception of the good life changes.

So the SMP and WMP concern themselves neither with instrumental desires nor with realizer desires. A good person will typically desire that they do the right thing, but they will desire that because the things they desire are actually the right thing to do, and they will (typically) know this. The principles say that the desires to do things that are actually right could be, or in the case of the latter two principles should be, explanatorily prior to the desire to do the right thing as such.

3.6. The Weak Motivation Principle (WMP)

3.6.1. Equilibrium

The WMP is restricted to equilibrium states. This restriction is there to deal with an important class of cases that Sigrún Svavarsdóttir (1999) discusses.

> [Smith argues that] the externalist account "re-describe[s] familiar psychological processes in ways that depart radically from the descriptions that we would ordinarily give of them" (M. Smith 1996, 180) ... Smith tells a story of a friend (let's call himMike) who has radically changed his moral view over the years from act-utilitarianism to a view that sanctions, in some instances, favoring family and friends, even when this cannot be given utilitarian justification. Since Mike is a moralist, his motivational dispositions have changed correspondingly ... I would like to offer an illustration of what sort of description externalists might give of Mike's mental states before, during, and after his two moral conversions. I venture the following speculation: Mike has always had some inclination to favor family and friends, but at one point he developed strong inhibitions against acting on these inclinations. These inhibitions were largely the result of being convinced that act-utilitarianism specifies the correct criterion for moral rightness. Having a strong desire to do the right thing and a rigid temperament, Mike quickly developed an avid interest in maximizing total happiness in the world, taking the interest of each person equally into account. In due time, his desire to maximize happiness actually started to dominate all other desires to the point that his friends thought of him as a utilitarian monster. But slowly doubts started to emerge as a result of exposure to arguments against utilitarianism. By and by Mike's conviction eroded and in the end he accepted a moral view according to which it is often right to be partial to family and friends, even when doing so cannot be given a utilitarian justification. At the same time, he came to see himself as a utilitarian monster, ever ready to sacrifice the interests of friends and family for the utilitarian project. Motivational dispositions he formerly took pride in having developed now became distasteful to him. However, since his desire to do the right thing has continued to be operative in his

> psyche, these dispositions are slowly eroding and the inhibitions on his inclinations to favor family and friends are undergoing radical change. They are gradually falling in line with his view of when it is right to give extra benefits to family and friends. (Svavarsdóttir 1999, 208–10)

Smith had argued that it is always a bad thing to be moved by the desire to do the right thing, as such. Svavarsdóttir's reply here is that this isn't bad at the very moment of major change in one's moral outlook. (Since this was the very example that Smith used against the motivational externalist, such examples were rather relevant to her debate with Smith.) Adopting a moral theory wholeheartedly requires adjusting one's motivations to align with it. But this need not be an instantaneous process; it can take time and effort. And the motivation to engage in this process of adjustment may come from a desire to do the right thing.

The defender of the WMP can concede all this. What the defender says is that Mike, in Svavarsdóttir's example, is not in equilibrium. What do we mean here by being in equilibrium?

For current purposes, it means having fairly settled moral views, and having had enough time and space since one's views became settled to make suitable adjustments in the rest of one's mind. Equilibrium requires the absence of felt pressure to change one's desires in light of changes to one's moral outlook.

Here are two cases that I take to not be in equilibrium, in the sense relevant to the WMP.

- Our hero faces a choice between competing values, and is torn about how to resolve them. She does not know which value is stronger, and she either lacks a clear disposition to resolve the tension in one particular way, or has such a disposition but does not trust it.
- Our hero systematically does not do what they believe to be best, and is trying to change their attitudes and behaviour to conform to their beliefs about the good.

On the other hand, the following two cases are cases of equilibrium in the relevant sense, albeit highly imperfect equilibrium.

- Our hero does not do what they believe to be best, but they have learned to live with this, perhaps feeling guilty about the gap between their thoughts and their deeds.

- Our hero is disposed to act one way, but would change their disposition if the reasons for acting a different way, reasons they already possess, were made salient to them.

In all four cases, the person already possesses something like reasons to change. But what makes for being in disequilibrium is the feeling that things must and will change.

Our ultimate interest here is in cases where moral beliefs do or don't line up with action, but we can come up with mundane, non-moral, illustrations of each of them. Here's a (schematic) illustration of the fourth kind of case.

I have a particular route I usually use going from B to C. I have a different route I use going from A to C. That route goes via B, but it does not take the usual route I use from B to C. This can't be optimal; if there is a best way to get from B to C, I should use it in parts of journeys as well as wholes. I could, nevertheless, be in equilibrium, even if a small suggestion (hey, why don't you do something different for the second part of the A-C route?) would push me to change my behaviour. The point is that equilibrium in the relevant sense just requires that the agent isn't trying to change, and isn't feeling pressure to change, even if they possess perfectly good reasons to change, and could easily be changed.

But in Svavarsdóttir's example, we do not have someone in equilibrium even in this weak sense. Mike wants to change his dispositions to line up with his moral theory, and he is making progress at this, but he still isn't there. The WMP does not deny that in cases like this, it is permissible to have goodness itself as a motivation.

3.6.2. Why Engage in Moral Reflection?

The following kind of consideration is sometimes advanced as a reason to be motivated by goodness as such. Sometimes people engage in practically directed moral reflection. That is, they think hard about what is the right thing to do, and the intended result of that thinking is that they do the thing they think is right. The most obvious analysis of what's going on in these cases is that the people involved want to do the right thing, and the point of engaging in reflection and acting on it is to bring it about that they do the right thing. And at least in cases where this leads to the thinker acting well, it seems this kind of moral reflection is a very good thing to engage in.

In the next section I'm going to say a lot more about this kind of case, because the SMP has to give a very different analysis of what is going on in moral reflection. But the defender of the WMP does not need to say much about these cases because they can simply endorse the 'obvious analysis'. The defender of the WMP can say that it is good, even optimal, to engage in moral reflection, motivated by the desire to do the right thing, when not in equilibrium.

The WMP is only making the following claim. When the storm is over and the seas are flat, a good person may be motivated by the things that make their actions right, not by the rightness itself. People who don't know what to do, and are torn between competing values, could not be a counterexample to such a principle.

3.6.3. The WMP and Two Kinds of Motivation Gaps

But why should we believe the WMP? I think the best reason is the simple intuition that Smith put forward: good people are motivated by things around them in the world, not by abstract notions of virtue and rightness. Another reason comes from reflection on fanatics like Robespierre and Saint-Just. But not everyone accepts those reasons. So let's look at a pair of cases that need explaining, and which the WMP can explain.

The first case is a petty crook who won't cross certain lines. In particular, while he'll steal anything from anyone, he won't engage in violence. This isn't just because he is scared of getting punished for violent acts. He has a kind of moral objection to violence. Perhaps speaking loosely, let's say that he has no respect for property rights, but a fitting and proper respect for rights involving bodily autonomy.

The thief's colleagues are planning a violent robbery. Feeling uncomfortable with this turn of events, the thief informs the police, who prevent the violence. This was a right and praiseworthy action by the thief. But what could make it right and praiseworthy? Not that he was trying to do the right thing - he's a thief who would have happily gone along with a non-violent plan to steal the goods. What makes his actions right and praiseworthy is that his motivation, prevention of violence against (relative) innocents, was good. There is nothing mysterious, and nothing wrong, with having this motivation without having a general motivation to be moral.

The second case is a person who has a desire to do what's right, but no underlying motivations. There are a couple of interesting variants of this case. Nomy

Arpaly (2003) spends some time on examples of 'misguided conscience'; people who want to do the right thing and are wrong about what it is. But we can also imagine someone who does want to do the right thing, and is broadly correct about what is right, but lacks any direct desire to do the thing that's actually right. Let's think about such a case for a bit.

Our protagonist, call him Rowly, was brought up well enough that he knows it is wrong to use violence to get things you want. And a desire to avoid wrongdoing was inculcated at a young age. So when Rowly wants a beer, but could only get one by punching someone, he declines to take the opportunity. But he is upset by this; he has no desire to avoid violence, or to avoid causing suffering, and wishes it was not wrong to punch someone to get a beer.

There is something deeply wrong with Rowly. We can see this by thinking about our interpretative practices. When someone says they did something because "it was the right thing to do", we do not normally interpret them as having no other-directed desires other than the desire to avoid wrong-doing. We do not normally think of such a person as being like Rowly. Someone who has to be taught what's right and wrong, and who has this belief as the only barrier stopping serious wrongdoing, is a deeply flawed human being. Even when people are too inarticulate to say what desires they have beyond a desire to do the right thing, we normally interpret this as inarticulateness, not a lack of respect for others, nor a lack of desire that others not suffer. This inarticulateness is not surprising; it's really hard to describe what makes actions right or wrong. But not wishing well for others is surprising; it's a serious character flaw.

So a desire to do the right thing is, in equilibrium, either unnecessary or insufficient. If one wants to prevent suffering to others, and acts on this, that's great, and it makes the desire to do the right thing unnecessary. If one lacks a desire to prevent (causing) suffering, then it is perhaps fortunate to have a desire to do the right thing, but that is insufficient for virtue.

Since a desire to do the right thing seems so useless, at least in equilibrium and in the presence of other good desires, it seems permissible to not have such a desire. And that's all WMP says.

3.6.4. Against Symmetry

I've argued so far that the WMP is true. I'm now going to argue that, assuming the WMP is true, there is an asymmetry between factual and moral uncertainty.

The role the WMP plays is to block one of three possible routes out of a problem facing the defender of symmetry.

We know that having the probability of some factual proposition move from 0% to 5% can (rationally) change behaviour. If I think the probability of rain is 0%, I don't have to check whether there is an umbrella in the car. If I think it is 5%, I will check the trunk to see the umbrella is still there before heading out. If symmetry holds, then changing the probability of a moral proposition from 0% to 5% should also change behaviour. And it is hard to see how that could happen.

I'm going to mostly assume here a broadly Humean picture of motivation: people do things that promote their desires assuming their beliefs are true. The relevant contrast here is with the view that beliefs, or at least belief-like states, can promote action without an underlying desire. So the Humean thinks I pack the umbrella because I believe it prevents me getting wet, and I have a desire to avoid getting wet, while the anti-Human thinks I pack it because I believe it prevents me getting wet, and I believe that it is good to avoid getting wet (or something similar).

I'm assuming the Humean view partially because it is implicit in our best formal models, partially because it seems intuitive, and partially because there are technical problems with the anti-Human view. David (Lewis 1988, 1996a) showed that the view that beliefs about the good played the role of values in expected value theory led to problems with updating mental states. Recently Jeffrey Sanford Russell and John Hawthorne (2016) have shown that these results rely on much weaker premises, and apply much more broadly, than a casual reading of Lewis's papers would suggest. Anyone who thinks that belief-like states alone can drive action has to adopt a rather implausible seeming picture of how beliefs are updated.

So I think rejecting belief-desire psychology is a high price to pay. But let's note it is one way out of the argument I'm about to give. I'll call it Option One for the symmetry defender.

If we don't take option one, then the symmetry defender must say which desires interact with a change in credence to produce a change in action. An obvious choice is to say that it is a desire to do the right thing. But that's blocked by the WMP. If symmetry is true, then there are times when a change in credence from 0% to 5% makes it compulsory to change actions. And it is not compulsory

to have a desire to do the right thing. So that won't work. For the record, Option Two for the symmetry defender is to reject the WMP, but that's also a bad move.

What the symmetry defender needs is to identify desires, other than desires to do the right thing, that can generate the action. These will be tricky to find. If someone thinks that it is 0% likely that doing X is wrong, then presumably it is completely rational to have no desire to avoid X, or avoid what X involves. So it looks like this route won't work either.

But that's too quick. All the symmetry defender needs is that after the change in credence, there is a desire that drives the change in action. Perhaps a change in credence could be correlated with a change in desires that produced, via orthodox belief-desire reasoning, the outcome the internalist wants.

But thinking there will always be such a change in desires is too much to hope for. Indeed, in some cases having such a change would be bad, as we can see using an example from Lara Buchak (2014).

Malai has a good friend, who she has known since childhood, and she values the friendship highly[4]. Then Malai learns that someone committed a horrible crime, and there is some very weak evidence that it was her friend. It's reasonable for Malai to have a slightly greater than zero credence that it was her friend who committed the crime, while not changing at all how much she values the friendship. Indeed, if the evidence is strong enough to move her credence, but not much more, it would be bad to have any other attitude. It's wrong to devalue friendships because you get some almost certainly misleading evidence about your friend. It's true the expected value of the friendship goes down when the evidence comes in, and if the friendship had only instrumental value, then that's a reason to devalue it. If Malai's only interest was in, say, getting to heaven, and she only valued the friendship insofar as she thought it likely it was a friendship with a good person, and that's the kind of thing that helps get you to heaven, then she should reduce how much she values the friendship. But most of us do not have quite that transactional an attitudes towards our friends or our friendships. Malai should have just as strong a desire to respect her friend and promote her friend's interests, and to respect and promote the friendship, as she had before getting the evidence. The evidence should not make her value the friendship less, and that's because friendships are intrinsically valuable, and how much

[4] I'm assuming throughout this paragraph that to value the friendship is a matter of having the right desires concerning the friend and the friendship, not having beliefs about the value of the friend or friendship.

something is intrinsically valued is not proportionate to one's credence that it is intrinsically valuable.

The same goes at the other end of the valuing scale. If one thinks that, for example, there is a 5% chance that purity is intrinsically valuable, it doesn't follow that one needs to (intrinsically) value purity at all. Nor does it follow that one needs to be motivated, at all, by considerations of purity.

I'll call Option Three the rejection of all that's been said in the last three paragraphs, and the insistence that changes in moral credences must occasion changes in desires. The examples involving Malai and involving purity make this option very unattractive.

Ultimately, I think this is the deepest problem for the symmetry view. Factual uncertainty changes our actions, and it does so rationally because it changes which factual uncertainty changes the expected value of different actions. For moral uncertainty to have the same effect, either we have to have a false view of the role of desire in action (Option One), or have to reject the WMP (Option Two), or have to adopt an implausible and unattractive view of how desires change when credences change (Option Three). None of these are correct, so symmetry fails.

3.7. The Strong Motivation Principle (SMP)

It is easy to imagine very good characters who are not motivated by the good as such; instead they are directly motivated by things that are actually good. Indeed, if one's motivations are fully in line with the good, it isn't clear what extra there is to be gained by also being motivated to be good. At worst, this motivation seems like either a distraction, or impermissibly self-centered. As Michael Smith puts it, people with this motivation "seem precious, overly concerned with the moral standing of their acts when they should instead be concerned with the features in virtue of which their acts have the moral standing that they have." (M. Smith 1996, 183)

There is something disturbing about a person who does not find the fact that a certain act is, say, a torture of a child to be sufficient motivation to not do it, and needs the extra motivation that it would be wrong. And the same goes for any other wrong act. Nothing is wrong as a matter of brute fact; there is always some explanation for why it is wrong. And that explanation always provides a

motivation that would prevent a good person from doing the action. Anyone who needs some further motivation is in some way deficient.

That is the intuitive argument for the SMP. And it seems to me compelling. But we can say more to motivate, and justify, the SMP. I'll start with a discussion of a central objection to the SMP; that it doesn't allow a special role for moral reflection. Then I'll discuss another reason to support the SMP; it avoids a certain kind of danger, one that we see manifest in history. And I'll close with a sketch of what a proponent of the SMP thinks the good person is like.

3.7.1. How to Explain Reflection

We typically think the following kind of activity is good. A person is faced with a difficult moral question, or with a question that she thought was easy, but which it turns out people she respects take a different view on. She reflects on what morality requires in such a situation. Upon coming to believe that morality requires of her something different than her current practices, she changes her behaviour to match with her new moral beliefs.

Such a character seems to pose a problem for the SMP. At first glance, it seems like a motivation to do good, or at least avoid doing bad, plays a central role. It is, apparently, the agent's change in her moral beliefs that triggers a change in action. And a change in a belief about what is X can only make a difference in action if X enters into one's motivational set in the right way. Since our agent seems to be a good person, it seems like good people should have thin moral motivations.[5]

My response to this kind of case will be very similar to what Arpaly and Schroeder (2014, 185ff) say about moral reflection. When our agent tries to figure out what morality requires of her, she won't start with highly abstract theorising. She will start with her concrete commitments concerning how she should engage with the world around her, and work out how those commitments apply to difficult or contested cases. As Michael Smith puts the point

[5] In the previous section I noted that the proponent of the WMP has an easy explanation of the appeal of moral reflection, since the agent who is motivated to engage in moral reflection is not in equilibrium. Since the SMP is not restricted to agents in equilibrium states, such an appeal will not work in defence of it.

> [N]ot only is it a platitude that rightness is a property that we can discover to be instantiated by engaging in rational argument, it is also a platitude that such arguments have a certain characteristic coherentist form. (M. Smith 1994, 40)

When good people use thin moral concepts in their reasoning, it is not because they are aiming at the good as such, but because these concepts are useful tools to use in sorting and clarifying their commitments, and making sure that they promote and respect the things they actually care about. We see this in other walks of life too. A competitor in a sporting event may steer their strategy towards moves that maximise expected returns. That's not because they care about expected returns; they want to win. It is because using the concept of an expected return is a good way to manage your thoughts when you want to think about how to win. And, in practice, this is often a very good way to manage your thoughts, so good strategists will use the concept. Similarly, it may turn out to be useful to use the concepts of goodness and rightness when trying to promote and respect the things that really matter, and so it isn't a surprise that we see good people using them.

3.7.2. Against Motivation by Morality

If moral concepts are useful tools for good people to use in promoting and respecting good aims, then we should expect that, like all tools, they have their limits. And indeed those limits are not hard to find. Moral reasoning is a kind of equilibrium reasoning. And equilibrium reasoning has clear strengths and weaknesses. There are cases when it is essential. Trying to work out the effect of a natural disaster on the market for widgets is practically impossible without doing at least some equilibrium reasoning. But there are also cases when it can go badly awry if not used extremely carefully, and in which very small errors in the inputs can lead to very large errors in the outputs. This is particularly the case when there are large feedback effects around. It is hard to use equilibrium reasoning to work out the effect of a rise in the price of labour, because changing the price of labour changes the demand curve for all goods, and hence raising the demand for labour. This isn't an insuperable modelling difficulty; but it means that it will take more than the back of a napkin to work out even approximately what will happen when the price of labour changes. Similarly, weather forecasting using equilibrium models is possible, but has to be done very carefully because very small errors in the initial inputs can push the modeller to an equilibrium that is far removed from reality.

We see the same problems when reasoning about morality. The method of reflective equilibrium, that characteristic coherentist form of reasoning, is the best method we've got for working out what is right and wrong. And it is very powerful. But it is an equilibrium method, and we are in a territory where there are very strong feedback effects. Whether one things X's treatment of Y is right or wrong will depend a lot on other moral judgments. If X is imprisoning Y, then that is probably very seriously wrong, unless Y has themselves done something seriously wrong, and X has been empowered (preferably by a good set of institutions) to deal with that kind of wrongdoing. Given there are this many feedback effects, we should expect that whether moral reflection leads people closer to, or away from, the truth is in part a function of how close they start to the moral truth. And this is, I think, what we see. To the extent moral reflection strikes us as a basically good practice, it is because we imagine it being used by people who have basically good motivations to start with. But in those cases moral reasoning will help smooth out the rough edges; it won't correct major faults.

And this suggests a problem with having morality itself as one of one's motivations: it is dangerous. Unless one starts with basically good motivations, thinking about the good and aiming for it could very well make things worse; perhaps catastrophically worse. We should acknowledge that in the hands of good people, moral reasoning can be a useful tool. The person who doesn't use that tool will almost certainly fail to optimise unless they have the sentiments of a saint. But someone whose aims include respect for others and their rights, freeing people from deprivation, promoting friendship and education, and being honest in their dealings, will usually act fairly well, even if they never engage in moral reflection. They may get the balance between these aims wrong from time to time, sometimes in ways that moral reflection would prevent. But they will typically avoid moral disaster. The person who aims for the good, as such, is more likely to land in disaster. One of the most dangerous things in the world is a wrongdoer with the courage of their convictions. Thinking about how and why equilibrium analyses can fail reinforces how dangerous this trap is.

But it's not just theory that tells us this is dangerous. The fanatic who thinks the individual is irrelevant, who will sacrifice any number of individuals to an idea, who will destroy villages in order to save them, is a recurring character in history. In some cases they are tragic figures; people who really did start out with praiseworthy aims but who refused to compromise when it turned out that those aims couldn't be realised without much suffering. And sometimes they are self-centred jerks, who feel empty unless they are trying to steer the whole world to their vision, whatever the costs. But what all of them teach us is that

aiming for the good, and just the good, can go terribly, horribly, wrong.

3.7.3. Back to Symmetry, and Moral Uncertainty

Let's turn away from these ideologues, and towards a positive picture of what a good but flawed person should look like. Our hero will mostly desire things that are actually valuable, and by and large desire them to the extent that they are actually valuable. They will have a well-functioning belief-desire psychology, so they will act so as to promote or respect those valuable things they desire. They will, from time to time, think about what is good and what is valuable, and form largely true beliefs about the good and the valuable. But since we are not supposing they are perfect, we will not assume these beliefs are inevitably true. And these moral beliefs, even the true ones, will not necessarily lead to much change in their action, because they don't connect up with any desire in the right kind of way. It is normal for a mismatch between desires and moral beliefs to lead to some unease, and to think that it might be wise to reform one's beliefs or one's desires. But depending on how deep the disagreement is, this reform program need not be a particularly high priority. And when it is carried out, there is no guarantee that the two will be brought into line by changing desires, as opposed to by changing beliefs. What there is a guarantee of is that if the moral beliefs conflict with other first order desires that the hero has, such as a desire that mass killings not happen, those other first order desires will play a powerful role in stopping the moral beliefs from taking control.

It is a thought almost as old as European philosophy that there is a good analogy between the well functioning polis and the well functioning mind. Although it is much less old, it is by now a venerable idea that the well functioning polis includes a separation of powers. And one of the virtues of such a separation of powers is that it limits the damage that can be done by a sudden swing in opinion among the powers that be. This is not a panacea; some states are rotten to the core, and no amount of institutional design will help. But it will prevent, or at least moderate, certain kinds of wrong. To put it in late 18th Century terms, the Alien and Sedition Acts were bad; the Reign of Terror was worse. It's worth thinking about what checks and balances in moral psychology would be, and more generally what a Madisonian moral psychology would look like.

My best guess is that competing desires, such as desires to promote welfare and alleviate suffering, and desires to keep promises and respect rights, are the appropriate kinds of balance to each other. But for current purposes it doesn't matter exactly how one ought implement checks and balances, only that it is good that

there are some. Because if moral uncertainty should be treated the same way as factual uncertainty, then there will be no checks and balances at all. When we firmly believe that some fact is true, then the thing to do is simply act as if that's true. We only hedge against the possibility that something is false when there is a possibility that it is false; not when we are certain that it is true. The symmetry view says that we should do the same with moral (un)certainty. But if that's the case, then there is no space for any check or balance on our moral views at all; when we are certain of them, they are guiding. That is wrong, and dangerous, so the symmetry view is also wrong.

Sometimes good people get the moral facts wrong. Perhaps they get bad advice, or bad evidence. Perhaps they start just a little wrong and equilibrium reasoning takes them to a place that is very wrong. When that happens, they have mechanisms to stop them acting seriously wrongly. I've been arguing that the moral mistakes shouldn't have any direct effect on action, because they won't aim at the good. But as I've noted already, I don't need anything that strong for the main argument of this book. What I need is that there should be some other forces that prevent action from lining up perfectly with moral belief when moral belief is seriously mistaken. A natural suggestion is that desires for things that are actually good can be that force. But even if that suggestion is wrong, as long as there should be some other force, then the symmetry claim fails.

3.8. Motivation Through Thick and Thin

In this section I'm going to run through some interesting test cases for WMP and SMP. I have two aims here. First, I want to strengthen the case for WMP. Second, I want to raise some cases that are useful intuition checks for testing the plausibility of the SMP. I know from talking to many people about the cases that I have different views about them to most people. So while I think the cases are evidence for a fairly strong version of the SMP, I know that they won't strike many people that way. Still, I hope the cases are useful ones for thinking about what's at issue in debating the SMP, and in particular thinking about how we should interpret the phrase 'moderately thin' in it if we want the principle to be plausible. But let's start with a case purely about maximally thin moral properties.

Milan is torn between two theories, and two actions. He gives some credence to an agent-neutral form of consequentialism, and some credence to a Kantian ethical theory. And he is torn between making a moderate donation to charity, one of 3% of his income, and a much larger donation to charity, one of 30% of

his income (which is all he can reasonably afford). He thinks that if the Kantian theory is true, then he isn't obliged to give more than 3%, and really doesn't want to give any more than he has to give. But he knows that if the consequentialist theory is true, then he is obliged to give (at least) the much larger amount.

Now Milan thinks most of the arguments favour the Kantian theory. But he has one remaining worry. He knows that the theory relies on having a workable notion of what it is for different people to do the same thing. And he worries that we don't have such a workable notion, for reasons familiar from philosophy (Goodman 1955) and game theory (Cho and Kreps 1987). So he sets out to do some philosophical research, reading about work on the notion of same action, and thinking about whether any such notion can generate a version of the categorical imperative that agrees with its intuitive content, and is not trivial. As often happens when working through a philosophical problem, his views on which side is stronger changes frequently. All the time, he has a web browser open getting ready to hit send on a donation. And as he changes his mind on whether the grue paradox ultimately defeats Kant's theory, he keeps adding and deleting a final zero from the amount in the box saying how much he will donate.

The WMP says that moral agents are not obliged to be like Milan. They don't have to have their charitable actions be sensitive to their beliefs about technical problems for Kantian ethics. It is, I think, reasonable to have one's credence in the correctness of Kantian ethics turn on beliefs about relatively technical problems. (For what it's worth, I think the kind of problem Milan is worrying about is a genuine problem for some kinds of Kantian theory, particularly those that think the formality of the theory is an important virtue of it.) But an agent who is being epistemically reasonable need not have their actions be sensitive to their technical worries. And that's because the agent need not be motivated by rightness as such.

If we change the case a little, we get an interesting test for SMP. Unlike Milan, Torin is convinced that some kind of Kantian theory is true. He also thinks there are technical problems with getting the formulation of the categorical imperative right. But he also thinks, sensibly enough, that these kind of technical problems are challenges, not reasons to reject the theory. Still, the way to solve the challenge will be to formulate different versions of the categorical imperative, and test them. And these different versions will have different consequences for which actions are required in certain circumstances. Is it reasonable for Torin to be differently motivated when he changes his views about which is quite the right formulation of the categorical imperative? I don't feel that it is, but I can imagine that different people have different views here.

A slightly more natural case seems even trickier to come to a firm judgment about.Florentina is trying to figure out what to do in a case where there are competing reasons in favour of two incompatible actions. She feels rather torn, but can't settle on a particular choice. Then she notices something: one of the choices, but not the other, is incompatible with the categorical imperative. Is it reasonable for her to be now more motivated to do the one that is consistent? I think this is a somewhat strange mindset, but I suspect many will disagree. What makes this case tricky is that we have to distinguish two situations that are rather hard to keep apart. We aren't interested in the case where Florentina sees that a choice is incompatible with the categorical imperative, and by seeing this sees that she had been overvaluing its strengths or undervaluing its weaknesses. Rather, we are interested in the case where this fact about the categorical imperative is itself a new motivation, alongside all the old motivations, to not do a particular action. To the extent I can keep a clear grip on the case, I think this is not a reasonable stance for Florentina to take. And that's why I think that it is wrong to be motivated by an action's compatibility or otherwise with the categorical imperative. What is reasonable is to see incompatibility with the categorical imperative as a reason for thinking there is something else wrong with the action, perhaps something we haven't yet seen.

Florentina's case is interesting even if you think that basing a whole moral theory around the categorical imperative is implausible. You can think that such a theory is surely wrong, but also think that Kant was nevertheless on to something important. Whether one could rationally will that everyone does X could be a factor in determining whether X is right or wrong, even if it is a long way from being a central factor. My default view in first-order ethics is a kind of muddy pluralism, which acknowledges that many distinct moral traditions have important insights into the nature of rightness and goodness, but which rejects any claim to comprehensiveness these theories may make. Florentina's case suggests that even if you have such a kind of pluralist view, you still could reject the view that conformity with the categorical imperative is a good motivation.

Let's move to some cases that seem a little easier. (I owe the following case to discussions with Scott Hershowitz.) Mercurius is a professor in a large university. As with most professorial positions, Mercurius has a fair amount of control over how much work he does. Some of his colleagues do more for the department than anyone could reasonably require, some do less than anyone could think was reasonable. Mercurius is a reasonable department citizen, handling a perfectly fair share of the workload, but only just as much as fairness requires. Today, as sometimes happens, a request comes around from the chair for volunteers for an

unexpected task. Mercurius does not find the task intrinsically interesting, but he knows that none of his colleagues will feel any differently. He knows he will feel a bit bad for whoever ends up shouldering the task, but will feel worse if it ends up being him. Still, he is worried he hasn't done his fair share of the work. This is wrong, as I said he has done enough, but it isn't an irrational belief since it is such a close call. So he volunteers, being motivated by a desire to do his fair share of the collective work.

This strikes me, and most people I've spoken about the case with, as a perfectly reasonable motivation. There is nothing objectionably fetishistic about being motivated to do one's share of a task one values. And Mercurius does value the good functioning of his department, and knows that it requires that the members collectively take on some unpleasant tasks. So he acquires a motivation to take on this particular unpleasant task.

It isn't easy to classify Mercurius's desire using the terminology we discussed in the previous section. He certainly doesn't have an intrinsic desire to do the unpleasant task. And it isn't strictly speaking an instrumental desire. We can imagine that Mercurius knows that one of the usual suspects, the people who already do more than their fair share, will take on this unpleasant task if no one else does. And we don't have to imagine that Mercurius values their time more than his. Nor is it quite right to say that Mercurius's desire to do this job is a realizer desire of his desire that the department runs well. After all, if he had just taken on a similar task the previous week, he would not desire to take on this one, although its relationship to the good functioning of the department would be unchanged. The best thing to say is that Mercurius has an intrinsic desire to do his fair share of collective projects that he has joined, and given his (false) beliefs about his past actions, this creates a realizer desire to do this unpleasant task.

So that puts an upper bound on the extension of 'moderately thin' in SMP. There isn't anything wrong with having a desire to do one's fair share, i.e., being motivated by properties like fairness. But on the other hand, thinking about these 'fair share' or 'good teammate' motivations helps explain some otherwise tricky cases. Indeed, my suspicion is that most intuitive counterexamples to the WMP, or even the SMP, can be helpfully thought of as cases where the agent has some independent motivation for joining a team or a project, and then a desire to be a good member of that team or project.

That's what I want to say about, for example, this case from Hallvard Lillehammer (1997).

> Consider next the case of the father who discovers that his son is a murderer, and who knows that if he does not go to the police the boy will get away with it, whereas if he does go to the police the boy will go to the gas-chamber. The father judges that it is right to go to the police, and does so. In this case it is not a platitude that a desire to do what is right, where this is read *de re*, is the mark of moral goodness. If what moves the father to inform on his son is a standing desire to do what is right, where this is read *de dicto*, then this could be as much of a saving grace as a moral failing. Why should it be an a priori demand that someone should have an underived desire to send his son to death? (Lillehammer 1997, 192)

A well functioning justice system is a very valuable thing to have. There is nothing at all fetishistic about desiring that one's state have such a system, and that it be maintained. Yet a well functioning justice system requires collective action, and this generates issues about whether one is doing one's fair share. As noted above, it can be reasonable, and not at all inconsistent with WMP, to desire to do one's fair share of a group project. Here the father who informs on his son should be motivated not be a desire to do what's right as such, but by a desire to do one's fair share of maintaining a good justice system.

If that's the right analysis of the case, then the father should be less motivated the less difference his informing will make to whether the state has a well functioning justice system. We see this already in Lillehammer's version of the case; the injustice of capital punishment is a reason for thinking that informing is not really a way of doing one's share in maintaining a system of justice. But similarly, if the family lives in a state where justice is very much the exception, it's reasonable to be less motivated to inform on one's son. By analogy, if tasks like the one Mercurius is considering routinely go undone, so there is no good functioning to maintain, that's a reason to be less motivated to take on this task.

Finally, consider a case about welfare, which has interesting lessons for moral motivations. Xue believes that human welfare is entirely constituted by health, happiness and friendship. And she is strongly motivated to promote her own health, happiness and friendships, which is natural enough given that belief. She is also motivated to help others–she is no moral monster–but for now we're just interested in her prudential reasoning.

Xue is told that bushwalking is good for your welfare, though she isn't told whether it makes you healthier, happier or have better friendships. But the source of this information is very reliable, so Xue forms a desire to do more

bushwalking. And this seems reasonable enough. Is this a case where Xue is motivated by welfare as such, and reasonably so?

I think it isn't. We have to distinguish three possible states.

1. Xue is motivated to do things that have the property *promote my health*, and is motivated to do things that have the property *promote my happiness*, and is motivated to do things that have the property *promote my friendships*.
2. Xue is motivated to do things that have the disjunctive property *either promote my health, or promote my happiness, or promote my friendships*.
3. Xue is motivated to do things that have the property *promote my welfare*.

Assuming fairly minimal coherence, we can't tell the difference between 1 and 2 by just looking at Xue's actions. Whether 1 or 2 were correct, she would do the same things in almost all circumstances. Perhaps she would say different things if the issue of whether she had disjunctive or non-disjunctive motivations arose in conversation. But we need not assume she has any interests in such a question, or even a pre-existing disposition as to how she would answer it. But that doesn't mean that there is no difference between the states. It is, in general, better practice to attribute non-disjunctive attitudes to agents rather than disjunctive ones (Lewis 1994; Weatherson 2013). So we should think that we are in state 1 rather than state 2.

Similarly, given her beliefs about the nature of welfare, there won't be much difference between the actions she is motivated to perform in state 1 and in state 3. So the fact that she responds to the information that bushwalking is good for her welfare by developing a desire for bushwalking is no evidence that we are in state 3. It might just be that we are in state 1. Since there is independent intutive reason to think it would be unreasonable for her to be in state 3, and her desire for bushwalking in this case is reasonable, we should think that we're actually in state 1. In general, we should prefer to attribute a plurality of underlying motivations to agents, rather than disjunctive motivations (as in state 2), or higher-order motivations (as in state 3).

3.9. Moller's Example

I'll end this chapter by discussing an analogy D. Moller (2011) offers to motivate something like symmetry.[6]

> Suppose Frank is the dean of a large medical school. Because his work often involves ethical complications touching on issues like medical experimentation and intellectual property, Frank has an ethical advisory committee consisting of 10 members that helps him make difficult decisions. One day Frank must decide whether to pursue important research for the company in one of two ways: plan A and plan B would both accomplish the necessary research, and seem to differ only to the trivial extent that plan A would involve slightly less paperwork for Frank. But then Frank consults the ethics committee, which tells him that although everyone on the committee is absolutely convinced that plan B is morally permissible, a significant minority - four of the members - feel that plan A is a moral catastrophe. So the majority of the committee thinks that the evidence favors believing that both plans are permissible, but a significant minority is confident that one of the plans would be a moral abomination, and there are practically no costs attached to avoiding that possibility. Let's assume that Frank himself cannot investigate the moral issues involved - doing so would involve neglecting his other responsibilities. Let's also assume that Frank generally trusts the members of the committee and has no special reason to disregard certain members' opinions. Suppose that Frank decides to go ahead with plan A, which creates slightly less paperwork for him, even though, as he acknowledges, there seems to be a pretty significant chance that enacting that plan will result in doing something very deeply wrong and he has a virtually cost-free alternative. (Moller 2011, 436)

The intuitions are supposed to be that this is a very bad thing for Frank to do, and that this illustrates that there's something very wrong with ignoring moral risk. But once we fill in the details of the case, this can't be the right diagnosis.

[6]Though note that Moller's own position is more moderate than the genuinely symmetric position; he thinks moral risk should play a role in reasoning, but not necessarily as strong as non–moral risk plays. In contrast, I'm advocating what he calls the "extreme view, [that] we never need to take moral risk into account; it is always permissible to take moral risks." (435).

The first thing to note is that there is something special about decision making as the head of an organization. Frank doesn't just have a duty to do what he thinks is best. He has a duty to reflect his school's policies and viewpoints. A dean is not a dictator, not even an enlightened, benevolent one. Not considering an advisory committee's report is bad practice qua dean of the medical school, whether or not Frank's own decisions should be guided by moral risk.

We aren't told whether A or B are moral catastrophes. If B is a moral catastrophe, and A isn't, there's something good about what Frank does. Of course, he does it for the wrong reasons, and that might undercut our admiration of him. But it does seem relevant to our assessment to know whether A or B are actually permissible.

Assuming that B is actually permissible, the most natural reading of the case is that Frank shouldn't do A. Or, at least, that he shouldn't do A for the reason he does. But that doesn't mean he should be sensitive to moral risk. Unless the four members who think that A is a moral catastrophe are crazy, there must be some non-moral facts that make A morally risky. If Frank doesn't know what those facts are, then he isn't just making a decision under moral risk, he's making a decision involving physical risk. And that's clearly a bad thing to do.

If Frank does know why the committee members think that the plan is a moral catastrophe, his action is worse. Authorising a particular kind of medical experimentation, when you know what effects it will have on people, and where intelligent people think this is morally impermissible, on the basis of convenience seems to show a striking lack of character and judgment. Even if Frank doesn't have the time to work through all the ins and outs of the case, it doesn't follow that it is permissible to make decisions based on convenience, rather than based on some (probably incomplete) assessment of the costs and benefits of the program. (I'll expand on this point in section 6.1, when I discuss in more detail what a normative externalist should say about hypocrisy.)

But having said all that, there's one variant of this case, perhaps somewhat implausible, where it doesn't seem that Frank should listen to the committee at all. Assume that both Frank and the committee have a fairly thick understanding of what's involved in doing A and B. They know which actions maximise expected utility, they know that which acts are consistent with the categorical imperative, they know which people affected by the acts would be entitled to complain about our performance, or non-performance, of each act, they know which acts are such that everyone could rationally will it to be true that everyone believes those acts to be morally permitted, and so on. What they disagree about is what

rightness and wrongness consist in. What's common knowledge between Frank, the majority and the minority is that both A and B pass all these tests, with one exception: A is not consistent with the categorical imperative. And the minority members of the committee are committed Kantians, who think that they have a response to the best recent anti-Kantian arguments.

It seems to me, intuitively, that this shouldn't matter one whit. I'm not resting the arguments of this book on the intuitiveness of my views. That's in part due to doubts about the usefulness of intuition, but more due to how unintuitive normative externalism often is. But it is worth noting how counterintuitive the opposing internalist view is in this extreme case. A moral agent making a practical deliberation simply won't care what the latest journal articles have been saying about the pros and cons of Kantianism. It's possible (though personally I doubt it), that learning of an action that it violates the categorical imperative would be relevant to one's motivations. It's not possible that learning that some people you admire think the categorical imperative is central to morality could change one's motivation to perform, or not perform, actions one knew all along violated the categorical imperative. At least that's not possible without falling into the bad kind of moral fetishism that Smith rightly decries.

So here's my general response to analogies of this kind, one that should not be surprising given the previous sections. Assuming the minority committee members are rational, either they know some facts about the impacts of A and B that Frank is unaware of, or they hold some philosophical theory that Frank doesn't. If it's the former, Frank should take their concerns into account; but that's not because he should be sensitive to moral risk, it's because he should be sensitive to non-moral risk. If it's the latter, Frank shouldn't take their concerns into account; that would be moral fetishism.

4. A Dilemma for Internalism

In the previous chapter I argued against the idea that we should treat factual uncertainty and normative uncertainty symmetrically. In this chapter I'll assume for the sake of the argument that the arguments of the previous chapter are unsuccessful. The upshot of that would be that we prefer theories that respect this symmetry. But this preference cannot be absolute. As with everything else in philosophy, we have to ask what the cost of satisfying this preference would be.

And in this chapter I'll argue that the costs are not worth paying. There are three kinds of theories that are possible. There are the externalist theories that I favour, which unqualifiedly approve of doing the right thing. There are theories that adopt an unqualified version of symmetry, treating all uncertainty the same way. I'll argue that such theories are implausibly subjective. And there are theories that adopt a half-hearted version of symmetry. I'll argue that these theories are under-motivated. There is no theoretical advantage, I'll argue, by incorporating a half-hearted symmetry principle. And there is much to be lost by giving up the idea that one should do the right thing.

The argument I'm offering here is based on a very similar argument that Miriam Schoenfield (2015) offers against various kinds of normative internalism in epistemology. The idea our arguments share is that the more subjective an internalism gets, the less plausible its verdicts about cases are, while the more objective it gets, the less well it is motivated by symmetry. Schoenfield primarily is interested in developing a problem for some forms of normative internalism in epistemology, but as we'll see, the same dilemma arises for internalism in ethics.

4.1. Six Forms of Internalism

The following schema can be converted into one of six internalist theses by picking one of the 3 options on the left and one of the 2 options on the right.

- Rightness/Praiseworthiness/Rationality is choosing an action with the highest credal/evidential expected goodness.

In every case 'goodness' is meant to be interpreted de dicto and not de re. That is, what has highest *credal expected goodness* is a function of the agent's beliefs (or more precisely her credences) in various hypotheses about goodness. And what has highest *evidential expected goodness* is a function of her evidence about is and is not good. If we interpret 'goodness' de re, then the principle is consistent with various forms of externalism; the de dicto interpretation is what makes these internalist theses.

The six theses we generate that way are all very strong. They all offer both necessary and sufficient conditions for an interesting concept. In the next two chapters, we'll look at internalist views that only offer necessary, or only offer sufficient, conditions for one of these. But it's helpful to start with the strong views to see what constraints there are on a viable internalism.

And I really want the six theses to be understood in an even stronger way. They should be understood to be explanatory in a right-to-left direction. So the view in question is not just that rightness (say) is co-extensive with maximising credal expected value, but that some act is right because it maximises credal expected goodness. This is, I think, implicit in the internalists that I'll cite below. And it makes sense given the idea that factual and normative uncertainty should be treated the same way. Orthodox decision theory doesn't just say that rational action is co-extensive with expected utility maximisation, it says that some act is rational because no alternative has higher expected utility.

It will help to have some abbreviations for the six theories. I'll use abbreviations for all five of the possible choices, and concatenate them to get abbreviations for the whole theory. I'll use Ri for rightness, Pr for praiseworthiness, Ra for rationality, C for credal and E for evidential. So, for instance, here are two theses one can express using this terminology.

- RiE - Rightness is doing the action with the highest evidential expected goodness.
- PrC - Praiseworthiness is doing the action with the highest credal expected goodness.

I've picked these because they are close to two theses endorsed by Michael Zimmerman (2008). They aren't exactly what he endorses; he leaves it open whether agents should be using expected value calculations, or some nearby variant. But they are nice, clean theories, and for that reason useful for theorising about. And

Zimmerman is hardly the only theorist to endorse something in the vicinity. Andrew Sepielli (2009) endorses something like RaC, and Michael (M. Smith 2006, 2009) endorses something like PrC and RaC.

The short version of this chapter is that the following three theses are both true and deeply problematic for any kind of internalism.

1. Both RiC and PrC theories make false claims about cases of what Nomy Arpaly (2003, 10) calls "inadvertent virtue" and "misguided conscience".
2. The E theories are unmotivated; they are a compromise between two extreme theories, but they inherit the vices and not the virtues of those extremes.
3. The Ra theories posit an asymmetry between cases of factual and normative uncertainty that undermines another kind of symmetry the internalist takes to be intuitive.

So none of the 6 theories are true. But more than that, the way in which the 6 theories collectively fail suggests that the problem won't be solved by adding epicycles, or weakening the theories to deal with hard cases. There is no version of normative internalism in ethics that is both motivated and plausible.

Sections 4.3–4.5 will deal with each of these theses in order. But first I need to say something about the assumptions behind the chapter. In particular, I need to say something about which possible theses are being set aside until the end of the chapter. And saying something about why we're setting various views aside will help position this chapter in the rest of the book.

4.2. Two Difficult Cases

There are four ways one could try to motivate normative internalism: by appeal to cases, by appeal to principles about coherence, by appeal to principles about guidance, and by appeal to symmetry. The first two are notably absent in the literature on normative internalism in ethics, though they will play a major role when we turn to epistemology.

There are, to be sure, plenty of arguments that talk about cases where agents have specified credences in theories T_1 or T_2, but typically, these arguments will not specify what T_1 and T_2 are. See, for example, Gustafsson and Torpman (2014)

and the papers cited therein, for instances of this phenomena.[1] I don't think these are really arguments from cases, since nothing like a case that we can have intuitions about is specified until we are told at least roughly what T_1 and T_2 are. If we were told that, for example, T_1 is Saint-Just's theory that the world has been empty since the Romans, and T_2 is Ayn Rand's version of egoism, we would have an example that we could have intuitions about.[2] Lockhart (2000) does include some case studies where he assigns credences to particular moral theories - including Rand's but not as it turns out Saint-Just's. But this isn't part of his defence of internalism, it's in the service of arguing from his internalist theory to various claims in applied ethics.

Now it isn't a bad thing that internalists don't argue from cases to theories. Indeed, there has been much criticism in the literature on philosophical methodology recently of philosophers' reliance on cases. (See Nagel (2013) for a discussion of, and reply to, some of that criticism.) But it does reduce how much we have to discuss here.

It will also be best to leave pure coherence based arguments until we get to epistemology. There is something intuitive about the following argument. It is incoherent to think that X is the unique right thing to do, but instead decide to do Y. Incoherence, in this sense, is a kind of irrationality. So rationality requires an internal connection between moral beliefs and action. Rather than discuss that argument directly, I'll just note that it is no more powerful than the following argument. It is incoherent to think that p is the unique conclusion supported by a body of evidence, but nevertheless believe q on the basis of that evidence. Incoherence, in this sense, is a kind of irrationality. So rationality requires an internal connection between epistemological beliefs and, well, beliefs. That looks like a pretty good argument at first glance too. Indeed, it is hard to see why we could accept the argument about moral coherence that I opened the paragraph with and not accept this argument about epistemological coherence. Now I'll deal with this epistemological argument at great length in part II of this book, and argue that it doesn't work, so I'll largely set the moral version of that argument aside for now.

But there is one version of the coherence argument that I want to more explicitly

[1] And, for what it's worth, in the papers I've seen so far citing Gustafsson and Torpman (2014), though that may change.

[2] I'm being flippant in reducing Saint-Just's moral and political theory to his aphorism about the Romans, but the details aren't really that important for what's going on here. See Williams (1995) for a more serious treatment of Saint-Just's worldview, and the earlier references on Robespierre for more details on Saint-Just's biography.

set aside. Consider a theory that accepts all three of the following principles. (See Markovits (2014) for a sophisticated version of the kind of theory I have in mind, but note that I'm simplifying a lot here to make a methodological point.)

- One should always do the right thing, and one should do the right thing in virtue of the right-making features of those actions, not in virtue of one's moral beliefs.
- Rationality requires that ones moral beliefs include all and only the true moral propositions.
- Immoral action is irrational.

Such a theory might agree with something like RaC. At the very least, it will say that rationality requires doing the action with the highest credal expected goodness. But that's because rationality requires both that one give credence 1 to the true claim about which action is good to perform, and rationality requires performing the action that is good to perform.

Is this theory internalist or externalist? I don't think it helps to try to classify it. Just note that I'm setting it aside. More generally, I'm setting aside theories that make moral omniscience the standard for moral rationality. Rational people can make mistakes; at the very least they can fail to believe some truths. That's true in science, it's true in everyday life, and it's true, I'm assuming, in ethics and epistemology.[3]

I discussed the guidance arguments earlier in the book, and argued that they only supported an implausibly subjectivist version of internalism. Not coincidentally, that's going to be similar to what I say in this chapter about the symmetry argument But you might think there is another way to block the argument from symmetry to internalism. This chapter and the last have been focussed on the following argument.

1. Expected utility theory provides the correct treatment of decision making under factual uncertainty.
2. Factual uncertainty and normative uncertainty should be treated symmetrically.

[3]This isn't an argument for this assumption, but perhaps a quick explanation for why the assumption seems plausible to me is in order. All arguments I've seen for the view that rationality requires moral omniscience have some kind of enkratic principle as a premise. And for reasons I will go over in Part II of the book, I don't think these enkratic principles are very plausible. Claire Field (forthcoming) has a very good critical discussion of the arguments for this assumption.

3. So some kind of internalist theory provides the correct treatment of decision making under moral uncertainty.

That's not valid, because a lot of the terms in it are rather vague. But I'm not going to dispute the inference here; if the premises are both true, then they will support some kind of theory that I want to reject.

I'm also going to assume, for now, that premise 1 of this argument is basically correct. And this is a substantive assumption. There is one very important moral theory that rejects premise 1 (under one important disambiguation of it). That's the traditional consequentialist theory that says that the moral status of an action is a function of the consequences it actually has (Sidgwick 1874; Smart 1961). I'm simply going to assume that's false for now, and come back to it at the end of the chapter. Note that I'm not assuming that modern consequentialist theories, like the decision-theoretic consequentialism Frank Jackson (1991) defends, are false. I'm just setting aside views on which factual uncertainty is irrelevant to the moral status of an action.

So to recap, we're making two large presuppositions at this stage of the dialectic. The defence of these presuppositions is largely in earlier chapters, but as noted above, some of it is to come. The presuppositions are:

1. The best argument for normative internalism is an argument from the symmetrical treatment of factual and normative uncertainty. This is an argument for a kind of internalism because (contra traditional consequentialism) factual uncertainty matters to the moral and rational status of actions.
2. Neither rationality nor morality requires moral omniscience, so if the morality or rationality of an action is sensitive to the actor's actual credence in moral propositions, or to the rational credence in those propositions given their evidence, then in some sense what they should do will differ from what the true (but unknown) moral or epistemological theory says they should do.

4.3. Inadvertent Virtue and Misguided Conscience

The next three sections will defend the three principles from the end of 4.1. So our aim here is to defend:

- Both RiC and PrC theories make false claims about cases of what Nomy Arpaly (2003, 10) calls "inadvertent virtue" and "misguided conscience".

Arpaly's paradigm of inadvertent virtue is Huck Finn, so we'll start with her description of his story.

> At a key point in the story, Huckleberry's best judgment tells him that he should not help Jim escape slavery but rather turn him in at the first available opportunity. Yet when a golden opportunity comes to turn Jim in, Huckleberry discovers that he just cannot do it and fails to do what he takes to be his duty, deciding as a result that, what with morality being so hard, he will just remain a bad boy (he does not, therefore, reform his views: at the time of his narrative, he still believes that the moral thing to do would have been to turn Jim in). If one only takes actions in accordance with deliberation, or the faculty of Reason or ego-syntonic actions [...], to be actions for which the agent can be morally praised, Huckleberry's action is reduced to the status accorded by Kant to acting on "mere inclination" or by Aristotle to acting on "natural virtue." He is no more morally praiseworthy for helping Jim than a good seeing-eye dog is praiseworthy for its helpful deeds. This is not, however, how Twain sees his character. Twain takes Huckleberry to be an ignorant boy whose decency and virtue exceed those of many older and more educated men, and his failure to turn Jim in is portrayed not as a mere lucky accident of temperament, a case of fortunate squeamishness, but as something quite different. Huckleberry's long acquaintance with Jim makes him gradually realize that Jim is a full-fledged human being, a realization that expresses itself, for example, in Huckleberry's finding himself, for the first time in his life, apologizing respectfully to a black man. While Huckleberry does not conceptualize his realization, it is this awareness of Jim's humanity that causes him to become emotionally incapable of turning Jim in. To the extent that this is Huckleberry's motive, Twain obviously sees him as praiseworthy in a way that he wouldn't be if he were merely acting out of some atavistic mechanism or if he were reluctant to turn Jim in out of a desire to spite Miss Watson, Jim's owner. Huckleberry Finn is not treated by his creator as if he were acting for a nonmoral motive, but rather as if he were acting for a moral motive–*without knowing* that it is a moral motive. (9–10)

Here are a few basic truths about Huckleberry's actions in helping Jim remain free.

1. Huckleberry does the right thing.
2. Huckleberry does not do the wrong thing.
3. Huckleberry is praiseworthy for helping Jim remain free.
4. Huckleberry is not blameworthy for helping Jim remain free.

If a philosophical theory rejects any of those four claims, it is wrong. Here are two more claims that I think are true, though I'm not going to rest any argumentative weight on them, since I suspect they will strike most readers as, at best, controversial.

5. Huckleberry's upbringing, and in particular the testimony from his parents, friends and teachers, provides strong evidence for the false moral theory that he in fact believes, namely that morality requires him to turn Jim in, and Huckleberry's relationship with Jim does not provide strong enough counter-evidence to make that belief irrational.
6. Huckleberry is rational, and not irrational, to help Jim to remain free.

If all of 1 through 6 are true, then all 6 of the theories we started with are false. Turning in Jim maximises both credal and evidential expected goodness. But helping Jim is right (1), praiseworthy (3) and rational (6). So all six theories are false.

The argument of the last paragraph relies heavily on 5 and 6 though. If 5 is false, then the case does not show any of the E forms to be false. And if 6 is false, the story does not show either of the Ra versions to be false. So without relying on 5 and 6, and I'm not going to rely on them, we can't argue against all forms of normative internalism using just Huckleberry Finn. But we can argue against some forms. Consider first RiC and PrC. The Huckleberry Finn case shows these to be simply false. Huck does the right thing, and is praiseworthy, although he clearly minimises credal expected goodness (at least relative to the live choices).

Huckleberry is a case of what Arpaly calls 'inadvertent virtue'. We can also put pressure on internalism by looking at cases of what she calls 'misguided conscience'. I'll use some cases described by Elizabeth E. Harman (2011), focussing on her examples that involve currently contested moral issues. (As Harman notes, if you don't find these examples forceful because you don't agree with the underlying moral theory, you could easily 'reverse' the cases to make a similar point.)

> Consider someone who believes abortion is wrong and who yells at women outside abortion clinics. It is wrong to yell at women outside abortion clinics: these women are already having a hard time and making their difficult decision more psychologically painful is wrong. But this person acts in a way that would be permissible if her moral views were true. Another example is someone who believes abortion is wrong and who kills an abortion doctor, in a part of the country where there is good reason to think that this doctor's death will reduce the number of abortions. This person believes that he ought to kill abortion doctors if doing so would reduce the number of abortions that would be performed. A third example is someone who believes homosexuality is wrong who organizes a campaign against the legalization of gay marriage. He believes he is doing something morally good in organizing the campaign; in fact, in working to further oppression, he is acting wrongly. (458)

As it stands, the various versions of the C theories say that these three actors are either acting rightly, or praiseworthily, or rationally. And again, the first two of these evaluations are wrong, at least if abortion and gay marriage really are morally permissible. Note that I'm not here claiming that the false moral beliefs involved are normatively irrelevant; it's consistent with what I say here that the characters in Harman's stories are blameless without being praiseworthy. I'm going to argue against that view in the next chapter, but I'll set it aside for now. What we need to focus on first is whether their mistaken moral belief suffices for their action being praiseworthy, and it does not.

4.4. Ethics and Epistemology

In the previous section we looked at arguments against C theories; theories that linked normative statuses to the agent's own credences. In this section we'll look at E theories, with the aim being to defend this principle.

- The E theories are unmotivated; they are a compromise between two extreme theories, but they inherit the vices and not the virtues of those extremes.

I'm going to start by making the case against this, that the E theories are in fact well motivated. That's partially because I think most internalists in philosophy prefer these to the C theories. And it's partially because the E theories are an

interesting attempt to solve a hard problem. But the problem they are trying to solve is really not solvable; and the attempt just inherits the vices of the positions it is trying to avoid without any offsetting virtues.

The debate will get very theoretical very quickly, so to try to keep things a little grounded I'll start with a fairly familiar kind of case. Zaina has been threatened by a group of determined pranksters. She is told, convincingly, that unless she pranks one innocent person, the group will prank that person and one hundred other people this week. But if she does perform the prank, the group will perform no pranks this week. And she knows that whatever happens this week will have no effect on how many pranks the group performs after this week. The prank in question is unpleasant for its victim; Zaina would not like to be the victim of such a prank. And while it might be mildly amusing for onlookers and perpetrators, Zaina knows that each performance of the prank makes the world worse.

What Zaina doesn't know is what the correct moral theory is. She has studied some philosophy as an undergraduate, and gives some credence to a consequentialist moral theory, according to which she should perform the prank so as to minimise prank performances, and the rest of her credence to a deontological theory, according to which it would be wrong of her to directly harm an innocent victim of her prank. And this is, we'll assume, a perfectly reasonable reaction to the moral evidence she has been presented. (If you don't believe this is possible, substitute some other theories in which you do think a thoughtful undergraduate could be unsure between after some kind of introductory philosophy course, and which recommend different actions in a particular puzzle case. It is a little unrealistic to think that Zaina could know that the truth is in one of these two places, and that will matter a bit below.)

Zaina doesn't know what she should do. But she also doesn't know what action will maximise expected goodness. She knows that according to the consequentialist theory, performing the prank maximises goodness. She knows that according to the deontological theory, not performing the prank maximises goodness. But she needs to know a lot more than that to work out what maximises expected goodness. She needs to fill in two variables in the following table.

	Consequentialist (Pr = p)	Deontologist (Pr = 1-p)
Perform Prank	-1	-v
Don't Perform Prank	-101	0

The expected value of not pranking is -101p. The expected value of pranking

is $-p - v(1-p)$. Figuring out which of these is larger requires solving two hard problems: exactly how likely is it that the consequentialist theory is true, and how do you put the violation of a deontological duty on the same scale as the difference between better and worse consequences.

The latter problem is very hard, and we'll come back to it in chapter 6. Ted Lockhart (2000) had a nice idea on how to make progress on it, but Andrew Sepielli (2009) shows that it doesn't work. Brian Hedden (2016b) uses the difficulty of this problem to argue against internalist theories generally. William MacAskill (2016) thinks that the problem is hard enough that we should respond by not trying to maximise the expected value of some random variable in cases of moral uncertainty, but instead using tools from social choice theory such as voting methods. I'm very sympathetic to MacAskill's approach, insofar as I think that conditional on us wanting an internalist theory of action under moral uncertainty, I think using tools from social choice theory is more promising than trying to find a value for v. But if we go down this route, we've given up the symmetry between moral and factual uncertainty, and as I argued at the start of this chapter, without that symmetry it is very hard to motivate internalism. So I'll assume that Zaina has to find out, or at least be sensitive to, the value of v.

Now the normative externalist has an easy thing to say about Zaina's case. If consequentialism is the true moral theory, then she should perform to prank to spare the other 100. If the deontological theory is true, then she should not perform the prank, since she should not commit such an immoral act. And that's all there is to say about the case. It might help Zaina to know what the right moral theory is, but it isn't necessary. If she performs the prank out of care for the welfare of the 100 people she is saving then, if consequentialism is true, she does the right thing for the right reasons. If she declines to perform the prank because it would disrespect the victim of the prank, then, if the deontological theory is true, she does the right thing for the right reason. Neither of the last two sentences require that Zaina know that she is doing the right thing or that her reasons are right - what's needed at most is conformity between her motivations and the right-making features of actions.

But the internalist tends to find this answer unsatisfactory for two reasons. The reasons tend to pull in opposite directions. The first reason is that it is in one respect too demanding. While it does not require Zaina to know something she has insufficient reason to believe, namely what the right thing to do here is, it does require her to be sensitive to some fact she is unaware of. That fact is, simply, what the right thing to do in this situation is. The second reason is that it is in a different respect too weak. Zaina could be massively incoherent, and the

externalist would find nothing wrong with her. Indeed, my preferred version of externalism says Zaina should be incoherent in some respects. It says that if true moral theory says that some factor is of no significance, then Zaina should give it no weight in her calculation, even though she thinks, and *should think* that there is a decent probability this factor is very morally important.[4] And many philosophers seem to find it extremely implausible that Zaina could be right, and rational, and praiseworthy, all without qualification, while there is a serious mismatch between her moral beliefs and her actions.

So let's try the opposite extreme, one suggested by our discussion of Descartes in chapter one. (Though what we start with will not be the view Descartes actually endorses.) What matters for morality is match between credences and action. So as long as Zaina does what she thinks is best, or perhaps what maximises expected goodness, she does the right thing. In that case she acts rightly, is praiseworthy, and is rational. While she needs to find values for p and v, she gets them by introspecting her beliefs, not by hard looking into the external world.[5] And the hero of this internalist Cartesian story is bound to be coherent, at least in the sense of having their views about what to do match up with the actions that are within their control.

But such a theory says some odd things about a different character, Antoine, who was threatened by the pranksters just last week. Antoine believes, rightly, that such a threat is a terrible affront to his dignity as a free person. He further believes, wrongly, that the only appropriate response to such an affront is to kill everyone who makes the threat. Fortuitously, Antoine is as bad at figuring out how to kill as he is at figuring out who to kill, so no one gets hurt. But we shouldn't let this lucky break obscure the fact that what Antoine does is seriously wrong. And yet, the Cartesian internalist has a problem with this. Antoine does exactly what his conscience tells him to do. He is as resolute a person as one could look for. And he is a villain; someone to be loathed and avoided, not admired.

So there is an easy and natural way out of the problem Antoine poses. Indeed, it is one that is entailed by the rest of what Descartes says in philosophy. Antoine

[4] I try to offset the oddness of this result by adopting an extremely pluralist first-order moral theory, so very few things that are plausibly of moral significance turn out to be irrelevant. But I don't want my defence of normative externalism to turn on this pluralism.

[5] I'm setting aside, apart from in this footnote, two problems with this view. As Eric Schwitzgebel (2008) notes, we are often mistaken as what we believe. And thinking that Zaina's beliefs settle the value of v requires adopting a 'desire as belief' view that faces various technical problems (Lewis 1988, 1996b; J. S. Russell and Hawthorne 2016).

does believe that killing the pranksters is moral, but this belief is extremely irrational. What he should be guided by is not his actual worldview, which is abhorrent, but the moral evidence that he has. And while we can't say for sure how that evidence would resolve a problem like the pranksters, we know it would not endorse a massacre.

And this is, I think, a natural motivation for the E theories. There is something intuitively appealing about trying to find a middle way between the externalist view that requires people to do the right thing without saying what that is, and the kind of subjectivism that has nothing plausible to say about Antoine.

But there are still problems. Indeed, the problems with this kind of worldview were pointed out by Princess Elizabeth in her correspondence with Descartes. The core problem is that this 'way out' requires treating ethics and epistemology very differently, and there is no justification for this differential treatment.

Antoine doesn't just believe that the moral thing to do is to kill the pranksters. He believes that his evidence supports that conclusion. If we are to say that what he does is wrong in some respect, then we have to insist that this does not matter. What he should believe is a function of what the evidence actually supports, not what he thinks it supports.

But now a version of the demandingness objection returns with a vengeance. The internalist thought was that it really unfair to require Zaina to be sensitive to a fact that she does not know - namely whether a consequentialist or deontological moral theory is correct. The proposed response now requires that she be sensitive to two facts that she does not know, namely which values of p and v are best supported by her evidence. And worse than that, we have replaced one yes-no question with two quantitative questions. This does not feel like progress.

I've skated over a division between ways the internalist might require that Zaina be sensitive to her evidence. First, they might require that she have the beliefs that are best supported by the evidence, and then act as her beliefs maintain. This is the version of the view that requires a fairly strong form of normative externalism in epistemology. There is no guarantee that Zaina knows, or even could know, what the rational credence in consequentialism given her evidence actually is. So requiring her to have credences supported by her evidence is requiring her to follow a norm that she does not, and could not, know. And avoiding that was supposed to be a big payoff for internalism. So this way of defending the E theories seems unmotivated.

But alternatively, the internalist here might just say that Zaina has to be sensitive to her evidence, not that she must know what her evidence supports. As far as it goes, the externalist agrees with this point. The externalist view is that the following three things are in principle separable. (In every case, read 'believe' as meaning 'fully or partially believe'; this covers appropriate credences as well as appropriate full beliefs.)

1. What Zaina should do.
2. What Zaina should believe about what she should do.
3. What Zaina should believe about what she should believe about what she should do.

The E theories say that while 2 and 3 might come apart, there is a tight connection between 1 and 2. There's nothing incoherent about that. But it is rather hard to motivate. The following situation is possible. (And thinking through this situation is helpful for getting clear on just what the E theories are saying.)

The true moral theory is deontological, so true morality requires that Zaina not perform the prank. The rational values for p and v given Zaina's evidence are 0.2 and and 15. That is, violating a deontological norm is (according to Zaina's evidence) as bad as the consequentialist thinks letting 15 people be pranked is, but consequentialism is fairly unlikely to be true. So given Zaina's evidence, it maximises expected goodness to perform the prank. But Zaina's credal distribution over possible values of p and v is centred a little off those true values, centred on 0.15 and 20. And while this isn't right, her margin of error in assessing what her evidence supports concerning p and v is great enough that she can't know these are the wrong values. So given her credences, she thinks her evidence supports not performing the prank.

Given all that, what is the sense in which she should perform the prank, in which it would be more rational, or moral, or praiseworthy, to perform the prank? It's true that if she were better in some respect - in respect of having credences that actually tracked her evidence - then she would perform the prank. But if she were fully moral, she would not perform the prank. And if she maximised expected goodness given her perspective, she would perform the prank only if she were a little better epistemically, without being better morally. But what philosophical significance could that counterfactual have?

While the case is artificial, it fits a natural enough pattern. Someone makes a pair of mistakes. These are mistakes - they are irrational things to do - but they are perfectly understandable since the task in question is hard. Happily, the

mistakes offset, so the person ends up doing something they would do if they made neither mistake. But there is some other option that the person would take if they fixed one particular mistake. Does that fact mean that the 'other option' is something the person should do, or morally ought to do, or is praiseworthy for doing, or is rational for doing? It doesn't seem like it; it seems rather that all we can say about that option is this rather technical claim that it has only one of two salient vices.

And that's the pattern for the E theories in general. They are unhappy half-way houses. If we want people to follow standards that they cannot know in full detail, those standards may as well be the standards of true morality. If we don't want to require this of people, then what their evidence supports is not determinative of what we can demand of them. Just what the evidence supports is sometimes hidden too. But if we start being too permissive, we end up saying nicer things than we really want to say about Antoine. There are a lot of choice points here, but none of them lead to a viable version of normative internalism.

4.5. Rationality and Symmetry

In the previous two sections, I argued against five of the six theories we started with. All that is left is RaC, and that will be the focus of this section. We'll start with a case modelled on an argument that Nomy Arpaly gives in response to a theory of Michael Smith's (Arpaly 2003, 36–46), then turn to the difficulties the internalist could have in motivating RaC by symmetry considerations.

Think again about Huckleberry Finn. I said it was rational of Huckleberry to help his friend Jim. But that's obviously controversial. One might think it is rational for Huckleberry to do what he thinks is good or right. At least, doing what Huckleberry believes to be bad and wrong seems like a kind of irrationality. If so, Huckleberry is irrational, and this might lend some support to a theory like RaC.

The last sentence of the previous paragraph is a non-sequiter. If Huckleberry is rationally required to do what he thinks is good, it does follow that what he does is irrational. But it doesn't follow that turning Jim in would be rational, unless the requirement to do what one believes is good is the only rational requirement there is. And that's not true.

Let's leave Huckleberry for a second and think about a different character, Noah. Noah has a friend, Lachlan, who he is thinking of turning in as a runaway slave.

He firmly believes that it is a moral duty to turn in runaway slaves, and that Lachlan is such a runaway slave. But both of these beliefs are absurd. Noah lives in Australia in the early 21st century, and there is no slavery. And he has been exposed to compelling reasons at school to believe that slavery is a grave wrong, and that people who helped runaway slaves were moral heroes. But Noah has somehow formed the implausible beliefs he has, and is now deciding whether to act on them.

Noah is irrational. Noah's beliefs that Lachlan is a runaway slave, and that turning in runaway slaves is morally required, are both irrational. If Noah attempts to turn Lachlan in though, would that be rational? I doubt it. One might say that it would be irrational to not attempt to turn Lachlan in, given Noah's other beliefs. I rather doubt this too, but we don't have to resolve the question. Even if not attempting to turn Lachlan in would be irrational, it might also be the case that attempting would also be irrational. There is no rule that says anyone has a rational option in any situation, no matter how many irrational things they have done to create the situation. Turning Lachlan in is a manifestation of some extremely irrational beliefs; it is irrational.

As Arpaly points out, the only way to motivate the idea that Noah is rationally required to do what he believes is good is to impose very strong coherence constraints on rational thought and action. We have to say that rationality in action requires coherence between thought and deed, even when that clashes with doing what one's evidence supports. But turning Lachlan in would be bad even by the standards of coherence. Such an action would not cohere at all well with the mountains of evidence Noah has about slavery.

As I mentioned above, it is arguable that Noah's case is a rational dilemma. Perhaps Noah is irrational if he turns Lachlan in, since he does something that he has no evidence is a good thing to do, and he is irrational if he does not, since this actions do not cohere with his judgments. But even saying that Noah faces a rational dilemma does not help the internalist here. For if Noah is in a rational dilemma, that's still a way of saying that rationality does not line up with maximising expected goodness. After all, maximisation norms never, on their own, lead to dilemmas.

We will have much more to say about the possibility of dilemmas in cases like this in subsequent chapters. But perhaps it is useful to note here that even if Noah is in a dilemma, it is an extremely asymmetric one. Even if you think it is somewhat irrational to act against his best judgment and fail to turn Lachlan in, it is much more irrational to act on no evidence whatsoever, and actually turn

him in. So RaC doesn't even provide a way to track what is most rational, or least irrational.

So RaC is false. It isn't only one's belief in what is good that is relevant to what it is rational to do, one's basis for that belief matters as well. But there's another reason to be suspicious of the Ra versions of the principles. Recall Cressida, our example of a reckless driver. What she does is irrational. But that's not all that's true of her actions. What she does is blameworthy and wrong. If we want to accept the internalist's symmetry principle, we have to say that whatever is true of Cressida is true of Huckleberry Finn. Saying that Cressida and Huck are alike in one respect, namely that they are both irrational, isn't a way of endorsing symmetry.

In fact, thinking about the analogy with Cressida gives us a reason to think that Huckleberry really is rational in what he does. Assume, for reductio, that Huck's action is irrational, in the way that Cressida's driving is irrational. Cressida, of course, also acts wrongly. What is the relationship between the irrationality of Cressida's action, and its wrongness? If the irrationality wholly explains the wrongness, then the irrationality of Huck's action should 'explain' the wrongness of it. But that can't be right, since Huck's action isn't wrong. If the wrongness wholly explains the irrationality, then there is no argument from symmetry for thinking Huck's action is irrational, since there is no underlying wrongness to explain the irrationality. More likely, the wrongness of Cressida's driving and its irrationality are connected without one wholly explaining the other. Now the externalist has a simple explanation of that connection; Cressida's knowledge of the risks imposed by driving as she does explains both the irrationality and the wrongness. But that kind of explanation clearly does not generalise to Huck's case. Huck's evidence clearly does not explain both the wrongness and the irrationality of his action, since it isn't in fact wrong.

Put another way, the defender of RaC can either try and defend their view with a narrowly tailored symmetry thesis, one that just applies to rationality, or with the broader symmetry thesis that would apply to rightness and praiseworthiness too. If we use the broader symmetry thesis, then Cressida's and Huckleberry's actions are alike in rationality iff they are alike in rightness. But they are not alike in rightness, so they are not alike in rationality. So RaC fails, since they are clearly alike in rationality according to RaC. So the defender of RaC is forced to use a narrow symmetry thesis. But it is hard to see the motivation for the narrowly tailored thesis. Once we allow that people can be wrong about normative facts, and so can violate a norm while believing they are following it, it seems plausible

that one could be wrong about rationality norms, and so could be irrational while believing one is rational.

4.6. Conclusion

So far I have argued against six forms of internalism. As I noted at the start, internalism is not committed to the disjunction of these six forms, so there is yet no full argument against internalism. So it might be hoped that some form of internalism can be found that is not committed to any of the six theses that have so far been undermined.

Hopefully though, it should be clear why the argument so far generalises to other forms of internalism. If the motivations for internalism can be used to support anything, they can be used to support a kind of radical subjectivism. According to this radical subjectivism, rightness, praiseworthiness and rationality are all matters of conformity to one's own views. And conformity, in the relevant sense, is also to be understood subjectively; to conform to one's views in the relevant sense is to meet one's own standards for conformity. Such a view has to say implausible things about cases of misguided conscience like Antoine, so can be seen to be false.

This completes the arc of chapters 2 through 4. In chapter 2 I discussed some reasons for thinking that our theory should treat moral uncertainty the same way that it treats factual uncertainty, and how this idea has motivated a number of recent versions of normative internalism about ethics. In chapter 3, I argued that this symmetry idea was not as intuitively plausible as it first seemed, and that there were in principle reasons to think that moral uncertainty, and constitutive uncertainty more generally, should be treated differently to the way we treat factual uncertainty. In this chapter, I argued that even if those arguments worked, and a symmetric treatment of factual and moral uncertainty is a theoretical desideratum, we should reject symmetry because it leads to implausible subjectivism. The only way to really respect symmetry is to have a radical subjectivism, and that is implausible.

This gets to the heart of what I find unsettling about internalism. We start out with three classes of facts:

- Moral facts, e.g., genocide is wrong.

- Epistemic facts, e.g., it is irrational, given current evidence, to have a low credence that carbon emissions from human activity are causing global warming.
- Coherence facts, e.g., it is incoherent to prefer A to B, and D to C in the main example in Allais (1953).

It is easy to feel that one should have something to say about agents who are unaware of all the moral facts. And that can push one towards a theory where the moral facts themselves don't play a substantial role in evaluating agents, rather something that is more accessible plays that role. But what could that be? If we say it is evidential probabilities of moral claims, then we are left saying that some facts that are beyond some agents' ken, i.e., facts about what is evidence for what, are evaluatively significant. Moreover, this kind of view will have strange things to say about cases of inadvertent virtue in agents whose credences track their evidence. So we might want to say something else. If we say it is not evidence but credence that matters, we are left saying that our most important criteria of evaluation turn solely on the agent's coherence. And again, we can ask whether by 'coherence' here we mean actual coherence, or coherence as it strikes the agent. Facts about coherence are not obvious. It is incoherent to believe the naive comprehension axiom. It is incoherent to have the usual preferences in the Allais paradox. Some people think it is incoherent to will something that one could not will to be universally endorsed. Some people think it is incoherent to believe there are discontinuous functions on the reals. If we judge agents by how well their actions, beliefs and evidence actually cohere, then we are judging them by a standard that could well be beyond their knowledge. If we judge agents by how well the think their actions, beliefs and evidence cohere, we'll be back to saying that Antoine is a hero. Assuming we want to avoid that, we have to apply some standards beyond what the agent accepts, and probably beyond what they could rationally accept.

The business we're in here is trying to work out how to evaluate agents and their actions. To evaluate is to impose a standard on the agent, one that they may not accept, and may even lack good reason to accept. That's the crucial externalist insight. We don't escape that conclusion by making the standard epistemic rather than moral. Agents can disagree with their evaluators about epistemic matters. And we don't escape that conclusion by making the standard simply one of internal coherence. Agents can disagree with their evaluators about what is and is not coherent. That the correct standards of coherence are arguably a priori knowable isn't relevant here; arguably the correct standards in ethics and epistemology are a priori knowable too. It is plausible that the correct standards of

coherence are somehow true in virtue of their form, but it isn't at all clear what the normative significance of that is. Disputes about whether there can be discontinuous functions or contingently existing objects turn on principles that are true (if true) in virtue of their form, but nothing follows from that about whether one could rationally have anything other than a firm true belief concerning the correct resolution of such a dispute.

It is natural to think that we should try to find something relatively easy to use as our initial evaluation of agents. If one thought ethics is hard, but epistemology is easy, it would be natural to think that we should use epistemic considerations as our starting point. But epistemology isn't easy. Or, at least, it isn't the case that all epistemic questions are easier than all ethical questions. If one thought coherence questions were easy while ethical questions were hard, it would be natural to think that we should use coherence considerations as our starting point. But coherence questions aren't easy either. It's epistemically worse to believe that torturing babies is morally good than it is to believe naive comprehension. Normative internalism is a search for what Williamson (2000) calls a cognitive home, but no such home exists.

There is one loose end to tidy up. Perhaps there is another way to respect symmetry. We could respect symmetry by having a much more radical objectivism. If we agreed with classical consequentialists such as Sidgwick (1874) and Smart (1961) that the right thing to do is what produces the best consequences, irrespective of the agent's evidence or beliefs, we could respect symmetry without getting Huckleberry Finn's case wrong. I'm not going to have anything original to say about this kind of consequentialism, but I wanted to briefly rehearse the reasons I don't think this is a good way to save symmetry. (For much more, see Slote (1992, Ch. 15).)

This kind of actualist consequentialism gets the case of Cressida the reckless driver wrong. And the moves that consequentialists make in response to Cressida's case do not seem particularly helpful in Jackson cases, as Jackson himself emphasises. (Jackson 1991) And actualist consequentialism combined with symmetry can't handle the cases of Prasad and Archie. That combination implies that the parents should have the same attitude towards their past actions, and they should not.

None of what I have to say about actualist consequentialism is at all original, which is why I've left it to the end. And of course this view is externalist, even more externalist in a sense than my own view. That's why the objections of this chapter do not really touch it. The reason the normative externalist is not forced

into actualist consequentialism is that symmetry fails, as was shown in the previous chapter. It's true that if the arguments of that chapter fail completely, then a new argument could open up against normative externalism, as follows.

1. If normative externalism is true, then actualist consequentialism is true.
2. Actualist consequentialism is not true.
3. So, normative externalism is not true.

But give what we saw in the previous chapter, we should already reject premise 1. And the arguments of this chapter, showing that symmetry will have one or other kind of implausible consequence, provides another reason to reject premise 1.

Most discussions of normative internalism in the ethics literature to date have revolved around symmetry. But there are considerations other than symmetry that may seem to motivate a variety of internalism, and in the next two chapters I'll discuss them.

5. Blame and Moral Ignorance

If an argument from premises concerning symmetry to a conclusion about internalism worked, we would get a very strong conclusion. It would turn out that true morality is irrelevant to our judgment of actions and persons. All we should use, when judging someone's actions, is the moral compass that they have. Or, perhaps, the moral compass that they should have given their evidence. As I have stressed though, the arguments for symmetry might undermine the availability of this fallback. It really looks like that if symmetry proves anything, it proves that the only moral standard (and indeed only epistemic standard) is (perceived) consistency with one's own values.

Put so baldly, it perhaps it isn't surprising that symmetry-based internalism as it was developed in the last three chapters ended up looking like a hopeless project. So it is time to look at alternative motivations for internalism, ones that suggest somewhat weaker forms of internalism. The views we looked at so far said that being true to one's own self (in one way or another) was both necessary and sufficient for the applicability of some key moral concept. Over the next two chapters, we'll look at views that drop one or other direction of that connection. So in this chapter, we'll look at views which say that conformity to one's own values is a sufficient condition for blamelessness. And in the next chapter, we'll look at views which say that conformity to one's own values is a necessary condition for avoiding all vices.

5.1. Does Moral Ignorance Excuse?

In recent work on moral responsibility, many philosophers have argued that that blameless ignorance of what's right and wrong is exculpatory. Something is exculpatory if it provides a full excuse; it makes an agent not blameworthy for a wrong action they perform. So slightly more precisely, what these philosophers have argued for is a version of the following view.

> Moral Ignorance Excuses (MIE)
> If agent S does act *X*, even if *X* is wrong, S is not blameworthy for this if:
>
> 1. S believes that *X* is not wrong; and
> 2. This belief of S's is not itself blameworthy; and
> 3. The belief is tied in the right way to the performance of *X*.

The second and third clauses are vague, and getting clear on whether there is a way of resolving the vagueness that makes the theory viable is going to be a big part of the story to follow. But the vagueness also makes it plausible to attribute MIE to a lot of philosophers. A classic statement of MIE, that I'll return to at some length below, is in "Reproach and Responsibility" by Cheshire Calhoun (1989). But the view has been more recently defended by Gideon Rosen (2003, 2004), Michael Zimmerman (2008) and Neil Levy (2009). And I plan to argue against all of these views.

Although I'll focus on philosophers such as Rosen and Zimmerman who have openly defended MIE, I've crafted the definition of MIE so that it is endorsed by many of their critics. In taking on MIE, then, I'm taking on a much broader range of philosophers than those who describe themselves as holding that moral ignorance excuses.

MIE is not a reductive account of blameworthiness; it uses the notion of blameworthy belief right there in the second clause. And through the 1990s and 2000s, much of the debate around MIE turned on how to understand that clause. Rosen and Zimmerman use MIE to argue for a very strong view. They think that mistaken moral beliefs are very rarely blameworthy, so they think MIE implies that people are rarely blameworthy. Or, at least, there are very few cases where we can be confident someone is blameworthy. This view is rejected by, for example, Alexander Guerrero (2007) and William FitzPatrick (2008). Looking back to the earlier debate, Michelle Moody-Adams (1994) similarly rejects some of the practical conclusions that Calhoun (1989) draws. But all of these rejections are accompanied by acceptance of something like MIE. The complaint these philosophers are making is not that MIE fails, but that philosophers have been too generous in their application of clause 2. I'm arguing for the stronger claim, that MIE itself fails.

In recent years more philosophers have adopted the more radical view that MIE itself is false. Elizabeth Harman (2011, 2015) has argued against the view that moral mistakes can be exculpatory. Since Harman thinks that moral mistakes are themselves blameworthy, her view is in a technical sense consistent wth MIE.

But that's only because she thinks that strictly speaking, it never applies. And the broader view on responsibility I'm adopting draws on work by Nomy Arpaly (2003), Angela A. M. Smith (2005) and Julia Markovits (2010).

Harman notes that the debate has been misnamed. When we talk about moral ignorance excusing, what we really mean is that moral *mistakes* might excuse. If someone is extremely confident that X is wrong, but not quite confident enough to know it, few philosophers would say that mental state is exculpatory when X is done. If they have a justified true belief that X is wrong, but don't strictly know this because their belief is in some other way defective, no one takes their lack of knowledge to be exculpatory. What matters are cases of moral mistake; cases where an agent firmly and reasonably has a moral belief that's simply false. Harman notes that this point is at least implicit in Guerrero's response to Rosen (Guerrero 2007), and a similar point is made by Rik Peels (2010).

In order to keep terminological consistency with most of the debate, while avoiding getting caught up on the point of the last paragraph, I'll make some more terminological stipulations. Say that a person is thoroughly ignorant of a truth p iff she believes $\neg p$. And then the live issue is whether thorough moral ignorance excuses. I'll assume that when other writers hypothesise that moral ignorance excuses, the term 'thorough' has been elided. And I will join them in this way of writing.

5.2. Why Believe MIE?

There are three main classes of arguments that have been given for MIE. The first is what I'll call the Argument from Symmetry, defended by Rosen (2003, 64) and Zimmerman (2008, 192).

1. Cases like Adelajdra's and Billie's show us that mistakes about matters of fact can excuse wrongdoing.
2. Moral mistakes and non-moral mistakes should be treated the same way.
3. So moral mistakes can excuse wrongdoing.

I've in effect already offered two responses to this argument. Adelajdra and Billie don't do anything wrong, so they don't need an excuse, so they aren't reasons to think that non-moral mistakes are excuses. And in general non-moral mistakes are not excuses. Borrowing some terms from jurisprudence, mistakes of fact are defences, not excuses; they are reasons to find someone not guilty, rather than

reasons to not punish them despite their guilt. And chapters 3 and 4 are long arguments for thinking that premise 2 of this argument is wrong.

The second is an argument from motivation, which we could put as follows.

1. It is good, or at least blameless, to do X because one thinks X is the thing to do.
2. If an action is blameworthy, this blame must be traceable to some stage that led to the production of the action.
3. So if the belief that X was the thing to do is blameless, then so is the performance of X.

The long argument that Michael Zimmerman gives for a version of MIE is, I think, a version of an argument from motivation (Zimmerman 2008, 175ff). And the argument plays a central role in Gideon Rosen's discussion. He describes a characterBonnie who is, as he puts it, an "unreconstructed selfish creep" (77) Bonnie cuts in front of a father waiting in the rain for a cab with his family, for no good reason other than she wanted to get uptown in more of a hurry. It turns out later that Bonnie has been suffering from a virus, and one effect of this virus is that she ceases to view considerations involving others as giving her reasons for action. But it did so, remarkably, in a way that left Bonnie in a relatively coherent state. She reflectively endorses her self-centred behaviour, and dismisses the importance of traditional moral considerations. Indeed, she apparently can hold her own in philosophical argumentation when confronted with the standard arguments against the kind of nihilism or egoism (it isn't exactly clear which) she now espouses. Rosen argues it is reasonable for her to take arguments for conventional morality seriously, but she does that, so to get her act well we will need her to do more.

> But is it reasonable to expect more? Here is Bonnie. She blamelessly thinks that she has most reason to steal the cab. What do you expect her to do? To set that judgment aside? To act on what she blamelessly takes to be the weaker reason? To expect this is to expect her to act unreasonably by her own lights. This is certainly a possibility, but is it fair to expect it or demand it? Is it reasonable to subject an agent to sanctions for failing to exhibit akrasia in this sense? When these questions are raised explicitly, the answers can seem self-evident. No, it is not reasonable to expect a person to do what she blamelessly thinks she has less reason to do. No, it's not fair to subject someone to sanctions for 'pursuing the apparent

> good' when it is clear that she is blameless for the good's appearing as it does. (Rosen 2003, 79–80)

Given what I've said about moral fetishism, I have to think premise 1 of the motivation argument fails. It is not good to be motivated by the good as such. What is good, or at least what is best, is to be motivated by that which is good. We expect Bonnie to be motivated by good reasons, even if she falsely takes them to be bad reasons. And premise 2 is, as Manuel Vargas (2005) argues, far from obvious. Maybe every blameworthy act is downstream from a move that is blameworthy in isolation, but it isn't obvious why we should assume this.[1]

Now I should note that there is an exception I noted in chapter 3 to the general rule that it is best to not act on the basis of thin moral beliefs. And that exception provides a possible way to defend something like MIE in a very narrow range of cases. I'll come back to this in the discussion below of Calhoun's view.

The third argument for MIE comes from cases where an agent acts from moral ignorance, and apparently it is intuitive that they are blameless. Many of these cases do not elicit anything like clear intuitions. And in the cases that do elicit relatively clear intuitions, it is a further step to say the correct explanation of that intuition is that the moral ignorance explains the blameworthiness.

For example, consider JoJo, as described by Susan Wolf (1987). JoJo is the son of Jo, a vicious dictator. Jo rose through the ranks to become dictator in a coup, and we aren't supposed to feel any hesitation in blaming him for his misdeeds. But JoJo was born in the palace, and raised to be ruler. He hasn't known any life other than the life of the vicious dictator, and has never been exposed to other moral systems. Many philosophers intuit that when Jo ascends to the throne, and continuous the family business of being vicious dictators, he is less blameworthy than Jo.

But to get to MIE, we need two stronger assumptions, neither of which we can get from raw intuition. One is that JoJo is not just less blameworthy than his father, but that he is blameless. The other is that JoJo is blameless because he is morally ignorant. I'm going to argue that neither of these extra assumptions is correct. When we return to JoJo at the end of the chapter, I will draw on some insightful things that Elinor Mason (2015) says about cases like JoJo's to argue that MIE is the wrong conclusion to draw cases from like his.

[1] When Guildernstern says "There must have been a moment at the beginning, where we could have said—no. Somehow we missed it." (Stoppard 1967/1994, 125) he is voicing something like premise 2. And it's not obvious that Guildernstern is right.

5.3. Chapter Plan

The argument here is going to be a little more roundabout than in other chapters, so let's have a road-map to see where we're going.

- In sections 4 and 5, I'll set out some very general features of blame that I'll be taking mostly for granted in the discussion that follows.
- Sections 6 through 9 will discuss the idea that moral ignorance only excuses if it is connected to action in the right way. I'll argue that most of the ways we might try to make this vague notion of 'connected in the right way' precise lead to implausible theories. The only version of MIE that survives being precisified is very weak.
- Sections 10 and 11 will discuss two kinds of cases that this weak version of MIE might still cover: wrong actions that are done habitually, and wrong actions by people in very different moral cultures. In each case, I'll argue that there are better explanations than MIE for why blame might be eliminated on reduced in these cases.
- Finally, I'll return to the picture of blame that results from these discussions, and go over how it handles a number of difficult cases.

5.4. Blame and Desire

The main aim of the chapter is to argue that MIE is false. It posits a necessary connection between (rationally) believing actions have a certain moral property, and actions actually having some similar property. Since I'm in the business of rejecting all such necessary connections, it's my job to argue against it.

I think MIE fails for a fairly systematic reason. It makes beliefs central to whether someone is blameworthy, but blameworthiness is a matter of having the wrong desires. Here I'm following a version of the view defended by Arpaly and Schroeder (2014). Blameworthy people either desire things that are bad, or fail to (sufficiently) desire things that are good. Actions are blameworthy (or praiseworthy) to the extent that they manifest bad (or good) desires to have. While I'm directly following Arpaly and Schroeder, what I say here owes a lot to the broader tradition on moral responsibility that traces back to Strawson (1962).

One might think that a desire-based view of blame would immediately rule out MIE: it says that beliefs are relevant to blame, but in fact only desires are relevant

to blame. And beliefs and desires are distinct existences, so beliefs can't be relevant to blame. But that's too quick. After all, which desires an action manifests is a property of the beliefs of the actor. So it could be that moral beliefs matter because they affect which desires are manifest by an action.

Here's one (implausible) way that moral beliefs could matter to blame. Assume that the good person has one and only one desire: to do the right thing. Then if a person thinks that what they are doing is right, it shows that they are manifesting the one and only desire that a good person has, so they are blameless. That is a way to make MIE compatible with the desire-based view of blame. But it's implausible twice over. For reasons we've discussed already, it's not true that the *only* thing a good person wants is to do what's right. And this view would say that acts of misguided conscience are not just blameless, they are positively praiseworthy, since they manifest the one and only good desire.

We don't need such an implausible view to get something like MIE though, within a broadly Strawsonian view on blame. Even if good people have more than one desire, we might suppose that one of their desires is to do the right thing. Assume someone has some bad desires, and perform an act that manifests those bad desires, but also desires to do the right thing, and this very action also manifests that desire. Then a plausible version of the desire-based view of blame is that their bad action will be less blameworthy than a similar bad action by someone who only has the bad desires. To that extent, the false moral belief will be something that reduces blameworthiness. And if we call anything that reduces blameworthiness an excuse, then the false moral belief will be an excuse. It won't be a full excuse, which is what most defenders of the MIE want, but it will be excusing.

There is another, even more plausible, way that false moral beliefs could be related to blame.[2] Assume that false moral beliefs are somehow connected to false beliefs about practical reasoning. ('Connected' here is deliberately vague, and the vagueness will matter in what follows.) People who make mistakes in or about practical reasoning tend to manifest different desires than we might have thought they did. And that could turn out to matter.

We can see this with non-moral examples. Abbott and Costello are each offered a deal. If they take the deal, it will gain them $10 straight away, and then lose $1 every day for the next 30 days. Both of them take the deal. But they do so for different reasons. Abbott has a really steep discount function. He values $10

[2] I only realised the importance of these cases in discussions with Claire Field about her work on blame and normative ignorance. I'll return to these cases in section 10.

now much more than the loss of $30 over the next month. Costello is practically irrational; the deal makes him worse off by his own lights, but he does not realise this. I've stipulated that Abbott and Costello value the deal differently, but in practice we can often detect these values without stipulation. Imagine we point out to Abbott and Costello that taking the deal will leave them $20 worse off at the end of the month. Abbott will say, "Who cares? I want the money right now." Costello will say, "Huh, I hadn't realised that. I wonder if I can back out of the deal." On a desire based view of blame, Abbott is to blame for his own misfortune when in a month's time he is $20 worse off than he might have been. But Costello is not blameworthy, since his desires are not defective.

In general, people who are practically irrational might be blameless for what seem like wrong acts, because the act does not reflect their underlying desires. This isn't exactly the same thing as saying that normative ignorance excuses, but it is very close. The following two groups are not identical as a matter of conceptual necessity, but they have a huge overlap.

- People whose actions do not reflect their desires.
- People who do not know what actions are best given their desires.

Indeed, many people who are in both groups would cease being in the first group as soon as they ceased being in the second. The view I'm going to be defending is that whether someone is in the second group does not directly matter to how blameworthy they are. That is, MIE is false. But whether someone is in the first group matters a lot, and whether someone is in the first group might in practice depend on whether they are in the second group. So normative ignorance will in many real world cases be indirectly relevant to blame.

5.5. Blame, Agents and Time

I'm not going to try to present a full theory of blameworthiness, and then derive results about ignorance and blameworthiness from it. But I will record with two important general points about blame that will matter in what follows.

The first is that it is agents, not actions or outcomes, which are the primary subjects of praise and blame. I will still say, and have already said, that agents can be blameworthy for actions. (Peter A. Graham (2014), in the course of offering a plausible general theory of blame, denies even this.) But it is the agent, not the act, that is the focus of blame.

The second point, which has not received sufficient attention in the recent literature, is that blameworthiness is time sensitive. It seems very bizarre to say that a particular action, performed at t_1, is wrong at t_2 but not wrong at t_3. Perhaps that is even contradictory. But it is certainly not contradictory to say that the agent of that action is blameworthy for the action at t_2 but not at t_3. Indeed, such claims are often true, as in the following case.

Glyn is a twelve year old boy. He steals Mehdi's expensive new jacket. Glyn does not need a new jacket, he is not suffering from any kind of duress or compulsion, and he knows it is wrong to steal. But he wants the jacket, so he steals it. At the time he steals it, he is blameworthy for the theft.

Fast forward forty years, and Glyn is now a middle aged man. He has not gone onto a life of crime. He is no moral saint, but an ordinary mostly moral-law-abiding member of society. It would be wrong to still blame him for the theft. Indeed, it is overdetermined that Glyn is no longer blameworthy. Typically, adults are not blameworthy for the wrongs committed by their juvenile selves. And typically, people are not blameworthy for the wrongs they committed in the distant past. We can test this by varying Glyn's case in different ways. If Glyn turns into a decent 19 year old, it seems wrong to blame him for the actions of his 12 year old self. And if he steals the jacket at 22, it seems wrong to blame his 52 year old self for the theft.

The law backs up many of these intuitions. Except for cases of severe wrongdoing, we typically give people a clean slate when they become adults. Records of juvenile wrongdoing are sealed, so as to prevent past misdeeds being held against someone. In the UK, this principle is taken further. The *Rehabilitation of Offenders Act 1974* makes it the case that after a certain length of time, even adult convictions for minor to moderately serious offences are *spent*. It can be defamatory to describe someone as a convicted criminal, if their conviction is spent. The law recognises that after a while, people are not responsible for the misdeeds of their earlier selves. More generally, whether someone is blameworthy for an action might change over time.

In some cases, I suspect this change of status can happen rather quickly. Change Glyn's case so that a few weeks later, he has a change of heart. He sheepishly returns the jacket to Mehdi, and apologises. And, crucially, Mehdi accepts the apology. Now Glyn is no longer blameworthy for the theft. He was blameworthy, but in a case where the misdeed was not too excessive, where only one person was harmed, and that person has accepted an apology, the period of moral responsibility has passed. Glyn was blameworthy for the theft, but he is no longer.

What's crucial is that blameworthiness can be time limited, not anything in particular I've said about apologies, or even about juvenile wrongdoing. We should reject the 'branding' model of blameworthiness, that once a person is blameworthy for something, they are branded with a moral cross, and must carry this mark for eternity. Rather, blameworthiness can ebb and, occasionally, flow.

This matters for one of the arguments Rosen gives concerning Bonnie. Bonnie is an "unreconstructed selfish creep" (Rosen 2003, 77), who nevertheless is internally coherent. So far, so bad. Bonnie seems like an appalling person, even if it is rather sad that she has become an appalling person. But a few weeks later, the virus wears off, and she regrets having the views that she previously had, and of course acting on them. Is she now blameworthy for the terrible things she did while suffering the virus?

I think one could go either way on this. But it seems that question is separate from the question of whether she was then blameworthy for what she did. If you think Bonnie should not now be blamed for what she did while suffering the virus, because you think that in some sense she isn't the same person as the one who committed those misdeeds, then your willingness to let Bonnie off the hook now isn't even evidence that what she did wasn't then blameworthy.

Rosen actually notes that time might matter to Bonnie's case, but dismisses this consideration too quickly. He writes,

> You may think that blame is no longer appropriate, not because the act was not blameworthy when it was committed, but rather because time has passed and it is time for you to let it go. The judgment that forgiveness is now mandatory is not the judgment that it was unfair to blame Bonnie in the first place. It is the judgment that further blame would be unfair given the severity of the transgression. Since we want to focus on whether the act was blameworthy when committed, we need to set this thought aside. So let's stipulate that the offence was recent enough and serious enough that if Bonnie was indeed responsible, you are not yet required to forgive her. (Rosen 2003, 81)

But the last point is exactly what can't be stipulated. It isn't just passage of time or forgiveness of victims that makes blameworthiness go away. Sufficient change of character can too. That's why it doesn't take too long for juvenile wrongdoing to be morally expunged. Commit the misdeed at the right time, and it might be legally expunged in a few days. Morality doesn't use the same hard cutoffs the

law uses, but the principle is the same. Bonnie's change of character is much quicker, but not completely unrealistic. (Compare the case discussed by Burns and Swerdlow (2003).) And we should have the same verdict; her actions were blameworthy, but she is no longer to blame for them.

5.6. Acting In Ignorance is No Excuse

The next three sections are on clause 3 of MIE. The clause is ambiguous, and on the most natural interpretations, MIE is clearly false. So we'll look at whether there is any interpretation that makes MIE plausibly true. It will take a while to cover the possibilities, but at the end the only version of MIE left viable will be one that is very weak.

In the *Nichomachean Ethics*, Aristotle distinguishes between acting in ignorance of the wrongness of one's actions, and acting from that ignorance. Getting clear on just what this means is not easy. But doing so is crucial to finding a version of MIE that is plausible.

Consider first an extremely contented carnivore. We'll assume she's in a world where meat-eating is wrong. And we'll assume she is ignorant of that fact, as she chews away happily on a hamburger. But this ignorance plays no role in bringing about her eating. She certainly does not think to herself "It's a good thing this is permissible," as she eats away. She is not disposed to order different foods on learning that meat-eating is wrong. She would not eat differently were she to have different views about meat-eating. She regards the coincidence between her wants and what is, by her lights, morally permissible as a happy but irrelevant accident. She eats a hamburger because she wants a hamburger, and that settles things as far as she is concerned.

The ignorance that our carnivore shows does not excuse her. It is true that she is ignorant of the wrongness of her action. But she doesn't eat because she is so ignorant. So it does not affect the moral status of her action. The general point is that moral ignorance that merely accompanies a wrongful act doesn't excuse the act. The ignorance must in some way make a difference to the act.

There are two natural ways to think that moral mistakes could be relevant to actions in ways that are excusing. First, the action might be counterfactually dependent on the mistake. If the agent wasn't making the mistake, they wouldn't have performed the action. Second, the action might be motivated by the mistake. That is, the reasons the agent had for the action might have included the

mistaken belief. In many cases, these two will go together. But this actually makes things tricky for the idea that action from ignorance can excuse. The next subsection will show that adding a condition that the action was counterfactually dependent on the mistake does not provide a sufficient condition for blamelessness. And the following subsection will show that if ignorance ever does excuse, it isn't necessary that the ignorance is motivating. Indeed, it is sometimes necessary that the ignorance is not motivating. The space of cases in which ignorance excuses is, if not empty, exceedingly small.

5.7. Against Counterfactual Interpretations of Acting From Ignorance

The mere presence of a blameless moral mistake does not excuse. It is a little more plausible to think that actions that are in some way traceable to a blameless moral mistake are excusable. Here's a version of MIE that makes that idea rigorous.

> For any agent S, proposition p and action X, if
>
> 1. S blamelessly believes p; and
> 2. p is false; and
> 3. If S had not believed p, S would not have done X, then
>
> S is not blameworthy for doing X, since her ignorance of X is an excuse.

This proposal won't work for a reason Gideon Rosen notes (Rosen 2003, 63n4). Pasco has read online that his football team has lost. The website he reads this on is, as he knows, extremely reliable. But the website is wrong on this occasion. Since Pasco knows the website is generally reliable, he is blameless for believing his team lost. Pasco reacts to the report of the loss by throwing a brick through his neighbour's window. He wouldn't have done this had his team won. So all three conditions are satisfied, and yet Pasco's ignorance of the result of the football match does not provide an excuse for the brick-throwing.

We need to at least supplement the simple theory. Rosen suggests the following fourth condition.

> If p had been true, then S's action would not have been blameworthy. (Rosen 2003, 63n4)

I'm not sure why Rosen uses 'blameworthy' here, rather than 'wrong'. It seems unintuitive to say that a false belief that would have offered a mere excuse if true could actually furnish an excuse. But I won't press the point, since it doesn't matter for the larger debate. Nothing like this condition can work. Indeed, it seems very unlikely that we can hold onto the idea that actions done from moral ignorance excuse, while understanding the concept of acting from moral ignorance in terms of conjunctions of counterfactuals.

To see this, add another assumption to the example: Pasco is a moral nihilist. That is, he thinks that nothing is good or bad, right or wrong, blameworthy or praiseworthy. This doesn't affect what he does very much (unless you're unfortunate enough to be stuck in a philosophical discussion with him). It certainly doesn't affect whether he reacts to bad football news by quietly cursing that overpaid forward, or by tossing bricks around. And assume that this belief in moral nihilism is blameless; it is a natural enough reaction to the strange diet of philosophical reading he has had.

Now let p be the proposition *Pasco's football team lost, and moral nihilism is true*. Pasco believes that. It is false; doubly so since both conjuncts are false. If he did not believe p, he would have not thrown the brick through his neighbour's window. I'm making an extra assumption here, but it's a plausible addition to the case. The assumption is that possible worlds in which either the website reports the results correctly, or Pasco reads some other website to get the football score, are much more like reality than the world where he sees the error of his nihilist thinking. That is, if he were to see that p is false, it would be because he saw the first conjunct is false, not because he saw the second conjunct is false. Finally, if p were true, what Pasco did would not be bad, or wrong, or blameworthy.

The last point is a little delicate, in a way that I don't think helps Rosen's case. Moral nihilism is necessarily false. False global moral theories are, typically, necessarily false. So evaluating counterfactuals about what would happen were one of them true require thinking about counterfactuals with necessarily false antecedents. Such counterfactuals are, to put it mildly, not well behaved. I'm a little inclined to think, following Lewis (1973) that they are all trivially true.[3]

[3]But wait! Haven't I been talking all book about examples in which some things are true that are, in fact, necessarily false? Yes, I have, but there's no contradiction here. I think, following Ichikawa and Jarvis (2009) that we should understand philosophical examples as little fictions. And, contra Lewis (1978), I think there are good reasons to not understand truth in fiction in terms of counterfactuals. It would take us too far afield to go into these reasons, but see Gendler (2000) for discussion of some of the relevant considerations. If I'm right about both of those claims, then there are non–trivial truths about about what is true in a necessarily false thought experiment, but not about what would happen if a necessary falsehood were true.

And I suspect this will make trouble for any attempt to spell out the idea of acting from ignorance in the way Rosen suggests. But set that aside, because the issues are very hard, and because we don't need to address them. Any theory of counterfactuals should say that it is true that if moral nihilism were true, then Pasco's action would not be bad, or wrong, or blameworthy. And that's all we need to make trouble for Rosen's view.

So on Rosen's view, Pasco's false belief in the conjunction *My football team lost and moral nihilism is true* excuses the brick throwing. And that's implausible. To see how implausible, note that the belief on its own that moral nihilism is true is not exculpatory. Imagine first a person just like Pasco, except that this person planned to throw the brick either in anger or celebration, whether the team lost or won. He shares Pasco's false belief, but he doesn't have an excuse. Next consider a person who is like Pasco except he has correct moral beliefs, and knows he is acting immorally when he throws the brick. He too has no excuse. It is only the strange combination of views and dispositions that Pasco has that are excusing. And that's very implausible, even if one thinks that false moral beliefs could in principle excuse.

One possible move that could be made here is to restrict the quantifier in our principle about excuse. Perhaps we should say that a false belief is excusing only if it is a false belief in a proposition about morality, and satisfies these conditions. But such a move would be undermotivated twice over. For one thing, a core motivation for MIE is the argument from symmetry, and making this move is to insist on a huge asymmetry. For another thing, we need some special story about why such a restricted theory should be true, and the literature is not exactly forthcoming with such stories. Indeed, if we had such a restriction in place, it isn't clear we would even need Rosen's fourth condition. But, as Rosen acknowledged, we do need such a condition to get a plausible thesis about mistake and excuse.

5.8. Against Motivational Interpretations of Acting From Ignorance

Still, there is a natural enough fix to MIE. Pasco's false moral belief, either on its own or in conjunction with false factual beliefs, doesn't excuse because it doesn't play the right kind of role in his deliberations. For a false belief to excuse, it isn't sufficient for an agent's actions to be counterfactually sensitive to the presence of the belief. Rather, the belief must play some kind of affirmative role in the agent's motivations, not just the kind of regulative role that is implied by counterfactuals

like the one Rosen uses. The action must not just be sensitive to the presence of the belief, but in some way brought about by the belief.

That's intuitively why Pasco's false belief in the conjunction p is not exculpatory. Although he would not have acted had he not believed p, his belief that p doesn't play any role in bringing about the wrong action. So adding a requirement that the ignorance be motivating avoids that counterexample. But it introduces new problems. I'm going to discuss two counterexamples to the new version of MIE.

First consider Gusto. Gusto normally has little interest in morality. But he is interested in girls, and right now he is interested in Irene. She says that she will only date him if he does nothing immoral for a week. So for this reason, and this reason only, he develops a keen interest in morality. Sadly, he gets some things wrong. So even though he thinks it is morally acceptable to break a particular promise he made to Oleg in the service of a greater good, it is not, and he acts immorally when he breaks the promise. Oleg can blame him for breaking the promise, despite Gusto's instrumental desire to act morally. It is very strange to think that his promise breaking ceases to be blameworthy merely because it is driven by his desire to date Irene.

For another example, consider Sebastian and Belle who are, blamelessly, committed consequentialists. That is, they do the actions they think will have the best consequences, understood in a completely neutral manner. They are also siblings. But there is a difference between them. When faced with any choice of any importance whatsoever, Sebastian will first think to himself, "What will produce the most utility?". Having convinced himself that a particular action is utility maximizing, he will perform that action just because it is utility maximizing. Belle has simply adjusted her values and dispositions in such a way that she sees actions in terms of their utility, and is directly motivated to do the thing that is in fact utility maximising.

One day, their mother is sick in hospital. It isn't life threatening, but it is a bit scary, and she would be helped by a visit from her children. But neither of them visit. They are both volunteering at a soup kitchen, and don't want to leave their posts. Sebastian deliberates about what to do. He thinks "If I leave, some people will go hungry. That will produce more disutility than my mother's sadness. And it is bad to produce more disutility, and I don't want to do what is bad. So I'll stay here." Belle simply is moved by the plight of the hungry in front of her, and stays without deliberating.

Assume, for the sake of the argument, that the beliefs Sebastian and Belle have are blameless. And assume that the impersonal consequentialism they believe is wrong - they should go to visit their sick mother. Finally, note that they would not have ignored their mother's needs had they not had their false belief in consequentialism. As Rosen's proposal stands, both of them are blameless for their action, since their false belief in consequentialism excuses. But Rosen's proposal is false, as the example of Pasco shows. And the natural way to fix it puts a gap between Sebastian and Belle. Since Sebastian's false belief in consequentialism does motivate his decision to stay, but Belle's false belief does not motivate her decision to stay, Sebastian has an excuse but Belle does not.

This is the wrong way around. Sebastian is worse than Belle. The kind of hypermoralised thinking that Sebastian engages in is exactly the kind of 'one thought too many' thinking that Bernard Williams (1981) accuses consequentialists of. I think, following Railton (1984), that Williams's complaint against consequentialism misses the mark. Belle is a perfectly good consequentialist, but can't be accused of having too many thoughts. But I do think Williams is right that having one thought too many is a bad thing,. And we should not reward Sebastian for having one thought too many by excusing his lack of filial piety. Even if you think Williams's idea in general is too strong, it seems extremely appropriate here. By pausing to deliberate, and check for his own moral rectitude, Sebastian helps nobody. Deliberation takes time, and it is time he could have spent helping others. Indeed, Belle does spend that time helping others. It seems extremely odd to say that she is blameworthy because she kept on serving food rather than stopping for a bit of a think then, quite predictably, gone on serving the food.

Perhaps we can avoid both kinds of counterexamples just described if we modify the idea that false moral belief can excuse even further. For a false moral belief to excuse it must:

- Be blamelessly held; and
- Be relied on in action guidance; and
- Be blamelessly relied on in action guidance.

Note that I say 'action guidance' not 'deliberation' here, because I take it we want to say that someone can be guided by their beliefs without using them in deliberation. When I descend from a balcony by the stairs rather than jumping over the railing, I'm guided by my belief that jumping over the railing will result in injury, even if I don't deliberate using that belief. Typically, I don't deliberate at all before descending via stairs rather than via jumping, so I don't use any beliefs in deliberation. Now given what we said about Pasco, we'll need some

notion of action guidance that is stronger than counterfactual dependence, and that will be no small challenge. I'll return to that problem in the next section, because first I need to note some points about the restriction to reliance that is blameless.

First, this restriction gives us at least a chance of getting the cases of Belle and Gusto right. A philosopher who thinks that false moral beliefs can excuse should, I suspect, say that Belle blamelessly relies on her consequentialism. She isn't directly motivated by it; indeed it feels rather forced to say she is motivated by it at all. She's motivated by the needs of her clients at the soup kitchen. But she relies, in some sense, on consequentialism. On the other hand, it is plausible that Gusto doesn't get off the hook so easily. It is, perhaps, blameworthy to have a merely instrumental motivation to act morally. So while Gusto's false moral beliefs may be blamelessly held, and may be relied on in action, they are not blamelessly relied on in action.

Second, it is a commonplace that blameworthy acts have to be traceable to something blameworthy. Indeed, something like this is a key premise in an important argument for MIE. (Though recall that I expressed some scepticism about that argument.) Now note that someone who did X being motivated by p, where p was both a false belief blamelessly held, and a completely terrible reason to X, would still be blameworthy. And a natural story about why that's true is that relying on an irrelevant consideration when deliberating about whether to do something wrong is blameworthy.

But once we remember Williams's point about too many thoughts, we should see that this is a very tight restriction. In a lot of cases, it is wrong to directly bring considerations of the morality of the action into one's deliberation. Rather, actions should be guided by the facts in virtue of which the action is right or wrong. A false belief about morality should generally be behaviourally inert, and so is not clear it can ever be blamelessly relied upon.

5.9. Adopting a Decision Procedure and Acting on It

But, says the objector, it is not always wrong to think about right and wrong and use this to guide one's actions. Indeed, this is what one should do when faced with novel, hard cases. The objector I'm imagining here is making a point similar to something Sigrun Svavarsdóttir (1999) says in response to Michael M. Smith (1994). Smith argued that good agents would never be motivated by right

and wrong as such, but things that made actions right and wrong. Svavarsdóttir argued, in effect, that this should hold only in equilibrium. (See, for instance, her example of Mike on page 209. I also discussed this example in section 3.5, making a somewhat related point.) When an agent first reaches a momentous moral decision, it is fine that they are moved to act by it. In the long run, they should be able to be moved by the forces behind the moral truth. That is, in the long run they should reach an equilibrium between their moral beliefs and their motivations; things they believe to be good should become directly motivating. But it is too much to require one's motivations to turn on a dime, the instant belief changes, especially if the decision is one where there are weighty interests on either side of the scale.

This doesn't make any trouble for the Sebastian and Belle case from the previous section, for we can easily add to the case that they have been consequentialists for long enough that they should have by now reached this kind of equilibrium. But it does mean that there could be some cases where someone actually is moved by a moral belief, and is not thereby blameworthy. It is easiest to see that happening in novel cases where there is a lot at stake, morally speaking. In some of these cases, we might think, virtue requires both careful moral deliberation, and perhaps even acting on the result of that deliberation in advance of one's motives lining up with one's resulting view of the good.

But it turns out there is a distinct problem these cases pose for the view that moral ignorance is exculpatory. The problem is one raised by Alex Guerrero (2007), though I'm going to put his point in a slightly different way. The worry is that defenders of the view that moral ignorance excuses haven't been sensitive enough to time. If any kind of moral mistake matters, it is a mistake at the time of action. But it is all too easy, when thinking about cases, to focus on mistakes at the time of belief formation. If there can be cases where a belief is blamelessly formed, but the persistence of that belief is blameworthy, these will come apart. And that's just what happens, Guerrero argues, in some of the cases that we've just said looked most promising for the view that moral ignorance excuses.

Consider the example, used by both Rosen and Guerrero, of the ancient slaveowner. Rosen says that such people were often blamelessly wrong about the morality of slaveowning. They were blameless because they simply absorbed the prevailing morality of the day. No one around them questioned whether slaveowning was right or wrong, so they were under no obligation to do so either.

But, says Guerrero, look at things from the slaveowner's perspective. He sees families being torn apart. He sees people being cast into chains and thrown into

dungeons. When faced with such appalling cruelty, it is callous in the extreme to not wonder for a moment whether this is all an acceptable way to treat people. Perhaps it is blameless to simply absorb moral standards in childhood the way one absorbs a language. But retaining those beliefs, not subjecting them to question when faced with the misery one sees every day in the institution of slavery, is a very different matter. (In general, slaveowning is a pretty terrible case for the proponents of the view that ignorance is exculpatory, as argued by by Michelle Moody-Adams (1994).)

Guerrero puts forward these considerations in service of what he calls 'moral contextualism'. I agree with his reasoning, and his conclusion, but not the name he gives to the view. What he defends isn't analogous to the epistemic contextualism of Cohen (1986), DeRose (1995) and Lewis (1996b), but to the interest-relative invariantism of Stanley (2005), and Fantl and McGrath (2009). The closest epistemological equivalents to his view came after his paper, in the theories developed by Ganson (2008) and Weatherson (2012) that make belief relative to the agent's interests, stakes and deliberations.

When an agent is abstractly deliberating the morality of slavery, or even mindlessly absorbing the prevailing wisdom, the stakes are not so high. But when they head to the auction block, or commission a slave-catching party, the stakes are about as high as can be. Taking a belief formed in such a low-stakes setting, and acting on it without further consideration in a high-stakes setting, is blameworthy. Not reconsidering the belief in light of the change in stakes is itself blameworthy. So even if the formation of the belief that slavery is permissible is blameless, the retention of it through the course of deciding to acquire and retain slaves, need not be.

I've set up Guerrero's argument in terms of interest-relative theories of belief. But his conclusion need not rest on anything quite so controversial. Ross and Schroeder (2014) object to these interest-relative theories of belief. One key problem they raise is that such theories make change of belief without change of evidence too easy. Ross and Schroeder propose instead that belief should be constituted by defeasible dispositions to use propositions in inquiry. In high-stakes settings, we retain the belief, but the disposition to use the proposition in inquiry is defeated. It should be clear this is no help to the proponent of the view that moral ignorance excuses. On Ross and Schroeder's view, the retention of the belief that slaveowning is permissible is not blameworthy. Indeed, it may be wrong to change that belief on the basis of familiar evidence. But when one is actually deciding to enslave other people, one should lose the disposition to act on the

belief. Just having the belief is no guarantee that one can, or should, use it. And in such a high stakes case, one should not.

So we have, after a long detour, an answer to some of the rhetorical questions Rosen posed earlier. Bonnie believes that she has most reason to steal the cab; what do we expect her to do? On the Ganson-Weatherson view, we expect her to lose the belief in light of the stakes. On the Ross and Schroeder view, we expect her to lose the disposition to act on the belief in light of the stakes. Either way, there's no excuse for simply harming others on the basis of a prior belief that one would be blameless in so doing.

5.10. Calhoun on Blame and Blameworthiness

The considerations raised so far suggest that there will be very few cases of wrongdoing that are excused for the reasons that Rosen and Zimmerman raise. The wrongdoing must be counterfactually sensitive to the mistaken belief, and be motivated, or at least guided, by the mistaken belief, and both of these things must be blameless, and the belief itself, both in formation and retention, must be blameless, along with the use of that belief in deliberation. And it turns out these exceptions to the excuse condition are complementary, so between them they cover a vast range of cases. Indeed, at this stage it would be reasonable to speculate that there are no cases at all that satisfy all of these constraints.

But while the constraints are tight, there are some interesting cases that might comply with all of them. These are the cases that are at the centre of the classic treatment of normative ignorance, Cheshire Calhoun's "Responsibility and Reproach". Calhoun's position is more complex than most of the contemporary views, and that complexity reflects a sensitivity to where the really tricky cases are.

Calhoun thinks that blameless ignorance can excuse. But she also thinks it is hard to be blamelessly ignorant of the wrongfulness of your actions when the society you're in knows they are wrong. Blameless ignorance will, in almost all cases, require social ignorance. (This theme is echoed in more recent work by Miranda Fricker (2010).) But in cases of social ignorance, if we want to bring about social change, we may have no alternative but to blame wrongdoers who we think, when engaged in philosophical reflection, are blameless for their wrongdoing. I'm not going to engage with that last point, as interesting as it is, save to

note that it might go some way to assuaging the intuitions of those who find the idea that ignorance can excuse highly counter-intuitive.

Following Calhoun, I'll spend some time on cases that are relevant to the way that structures of sexist oppression are maintained by actions that are hardly oppressive in themselves. These acts are thoughtless, but on their own they are almost harmless. The problem, of course, is that these actions are not done on their own. There are a lot of sexist actions that are much worse than thoughtless, but we'll set those aside for now. So don't focus for now on the pimps and the pornographers, or even on the fathers who go out of their way to provide more for their sons than their daughters. Instead focus on casual, everyday sexism of ordinary men in sexist societies. ('Microaggressions' in current terminology.) Calhoun certainly wants to say that the things these ordinary men do are wrong. Indeed, they are collectively extremely wrong, as they collectively maintain a structure of oppression. But, she says, the ordinary men involved are not blameworthy for their misdeeds. There are three grounds for that claim in her paper.

First, blaming everyone would make us massively revise our views of the virtue of many people around us.

> If we assume, as we often do, that only morally flawed individuals could act oppressively, the we will have to conclude that the number of morally flawed individuals is more vast than we had dreamed and includes individuals whom we would otherwise rank high on scales of moral virtue and goodwill. The oddity of this conclusion forces serious questions about the possibility of morally unflawed individuals committing serious wrongdoing. (Calhoun 1989, 389)

This conclusion shouldn't strike us as odd. The world is full of people who have good features alongside serious character flaws. Indeed, it is common in psychological research and in classic literature and in history to see basically good people easily led down a path of moral corruption. (Robespierre was one of the most morally decent people in the French Revolution until he wasn't.) If a theory implied that basically half the population fell under the description *largely good but with a big moral flaw* I wouldn't see that as a fatal flaw in the theory.

In the case of casual sexism, the argument that we couldn't conclude that everyone is flawed seems particularly strange.Let's say we don't want to blame the ordinary man for the part he plays in oppression by everyday acts like using demeaning terms like 'girls' to refer to women. Still, most of these ordinary men provided considerably more resources for their sons than their daughters. And

of those that did not, most were 'off the hook' solely because they didn't have both sons and daughters. And that unfair distribution, or at least disposition to distribute unfairly, isn't the kind of individually minor wrong that Calhoun wants to excuse. It isn't that surprising to think that most men in sexist societies are at least somewhat blameworthy.

Second, Calhoun notes that we should not simply assume that causing harm is "the same as being responsible for the harm" (Calhoun 1989, 392). This is surely right - and I want to come back to it below. It is important that we respect the conceptual difference between wrong and blameworthy action, and return to that distinction.

But the biggest consideration, one that runs through Calhoun's paper, is that there are cases where the wrongdoer has no reason to reconsider their false moral belief. Calhoun shows that many false moral beliefs are not be like that. If society disagrees with one's false moral beliefs, then one has frequent occasion to reconsider the belief. So the false moral beliefs one has no reason to reconsider will only be one's shared by the community. For many other false moral beliefs, the thing that makes actions wrong will be a reason to reconsider the belief at the moment of action. (Guerrero's response to Rosen turns on a similar point.) But maybe that isn't true for all false moral beliefs. Maybe it is only true if the wrongness rises above a certain threshold. Not every potential harm is a ground for reconsidering one's views.

Let's consider a very different kind of example to Calhoun's in order to see the possibility I have in mind. Assume that Inka lives in a city with a busy underground train system. It is possible to delay a train's departure from a station for a few seconds by standing in a doorway preventing a train leaving. Inka believes, as do most people around her, that it is acceptable to so delay a train in order to let a friend running for the train catch it. Indeed, this is widely taken to be a requirement of friendship. It is also a false belief. Costing the other 600 people on the train 10 seconds each of extra travel time is a very bad thing to do in order to save one friend the five minutes they would have to wait for the next train. It is, in Inka's world, as morally bad as trapping one person in a train for 100 minutes.[4]

Now Inka is in a position where she can help a friend catch a train by blocking the doorway, and letting her friend run the last few steps to catch it. She does just this. It's not true that what Inka should have done instead was stop and

[4] I think holding a train for a friend to catch it is also wrong in *our* world, for just this reason. But the argument doesn't turn on this assumption.

think about her action. Any deliberation and the moment for action will have passed. Even if Inka did have time to contemplate action, it isn't clear she has any reason to do so. There is no person who she is harming so severely that this harm makes it compulsory for her to reconsider her actions. After all, each of them is only being delayed in the train for a few seconds more, and people get delayed in subways for seconds all the time. And Inka presumably thinks that small delays like this are of no moral significance.

Putting these points together, we can sketch a way in which moral ignorance might excuse. Assume that all of the following conditions are met.

1. S is, as a matter of policy, disposed to do *X* in circumstances C.
2. S has this policy for the same reasons that she has the false moral belief that doing *X* in circumstances C is permissible. This means that she would (eventually) lose the policy if she lost the belief, and that the adoption and maintenance of the policy is part of a general disposition to adopt policies she regards as permissible.
3. It is commonly held in S's community that it is permissible to do *X* in circumstances C.
4. S is in circumstances C.
5. If S does *X* right now, no one will be greatly harmed.
6. S has no time or reason to reconsider her policy of doing *X* in C before it becomes time to act.
7. S does *X*.

None of the arguments I've offered so far refute the idea that S is less blameworthy in these cases than is a typical person who does *X*. Even without having necessary and sufficient conditions for 'acting from ignorance', it is plausible that S acts from ignorance and not just in ignorance. And it's not true that the situation S finds herself in calls for reflection and deliberation rather than action. So nothing I've said so far rules out S's actions being excused. I'm going to defend three claims about this case.

- We need to know more about S to know how much her situation excuses her behaviour.
- Even when S has a partial excuse, it isn't the fact that she is morally ignorant that excuses. But the moral ignorance is relevant; the reasons that she is morally ignorant will typically be the reasons that she is excused.
- In any case, there are no such things as full excuses for wrong action, so S couldn't have a full excuse for wrong action.

I'm going to defend the first and second claims in this section, and the third claim at the end of the chapter. The defences will turn on considerations arising from cases of practical irrationality. [5]

Imagine that you have the following views. You think that the conclusions of "Famine, Affluence and Morality" (Singer 1972) are basically correct. And you're a Strawsonian about moral responsibility. And you know of someone, call him Gloucester, who doesn't give a lot to charity. But you also know that Gloucester is disposed to massively increase his charitable giving were he to simply be presented with Singer's drowning child argument. If he were presented with that argument, he would have an "A Ha!" moment, and see that he was required to give much more to charity. What should you say about Gloucester's responsibility for his current insufficient charitable givng?

I think you should say that Gloucester is largely blameless. He would be completely blameless after he reads Singer and starts donating. But reading the paper does not change his fundamental desires. And the Strawsonian picture is that these fundamental desires are what makes him praiseworthy or blameworthy. So he isn't particularly blameworthy now.

Does this mean Gloucester is a case where moral ignorance is excusing? I don't think it does. Gloucester is practically irrational. Right now, before reading Singer, he values charitable donation over spending money on himself. But he doesn't act on those values. He also doesn't realise that those are his values. And both the failure to act and the failure to realise have a common cause - his practical irrationality.

Now Inka might be just like Gloucester. It might be that as soon as you present her the simple argument for why holding the train doors is wrong, she has an "A ha!" moment, and sees that what she is doing is wrong. That is, she sees that by her own lights, it was better to not hold the train door. And what matters is not that she sees this, but that it was true all along that she desires implied that train doors should not be held in these cases. She presumably also had other desires, inconsistent with those, that implied that the doors should be held open. And she is somewhat blameworthy for having those desires; but the existence of the correct desires mitigates her blameworthiness.

But Inka might be different. It might be that she needs to change her desires to change her actions. She might be like someone who gives more to charity after

[5] As I mentioned above when introducing the cases of Abbott and Costello, I only saw the importance of these cases in discussions with Claire Field. Her work in progress has a different, and interesting, take on the cases I'm about to describe.

visually seeing the suffering of the impoverished. Such people, I think, change their desires upon seeing the suffering. And they are blameworthy for not doing more in the first place. (At least if the conclusion of Singer (1972) is correct, and I'm not assuming it is for anything more than the duration of this argument.)

So that's my overall conclusion about the people who commit minor wrongs out of habit, while not believing they are wrong, and not having social pressure to change. Some of these people all along had desires that the wrongs not be committed. They were practically irrational in committing the wrongs. They have a partial excuse. They are also morally ignorant. But the moral ignorance does not explain the excuse. Rather something else, their practical irrationality, explains both the excuse and the ignorance. I don't want to take a stand on whether this is rejecting MIE, since it says ignorance never makes one have an excuse, or endorsing a weak version of MIE, since it says that ignorance goes along with having an excuse in some cases. That would require more careful parsing of the words of MIE-defenders than seems useful to do here.

Instead I'll turn to one other aspect of Calhoun's examples. Inka is a case of habitual minor wrongdoing, in cases where habitual rather than reflective action is called for. Now we'll spend a bit of time on wrongdoing that is culturally approved.

5.11. Moral Mistakes and Moral Strangers

Nothing I've said so far explains why JoJo is less blameworthy than his father. And many philosophers hold it to be very intuitive that he is, and that something like MIE explains why. My preferred account of JoJo relies on work by Elinor Mason (2015).

Mason argues thinks that the debate about moral ignorance has been oversimplified in a number of ways. She argues that there are really two kinds of blame, what she calls 'ordinary blame' and 'objective blame'. And MIE is completely wrong about objective blame. But it is correct for ordinary blame, provided we are careful to restrict clause 2 to full beliefs of the agent. This restriction excludes where the agent knows, or even suspects, deep down that they are doing something wrong. And she is much more willing than Rosen or Zimmerman to say that even when someone does fully believe that what they are doing is right, this could be due to a blameworthy kind of motivated reasoning. But all that said, she does think that a suitable version of MIE could hold for ordinary blame. Here

is an important part of what she means by 'ordinary blame', and why she thinks MIE is correct about it.

> Normally, we blame each other for what we deliberately do. And if we find out that some piece of behavior was not deliberate, we let the agent off the hook. This is ordinary everyday blameworthiness. Ordinary blameworthiness is based on subjective wrongdoing. When ordinary people behave badly, they are usually, at some level (and this need not be the fully conscious level that Rosen and Zimmerman require), aware that they are doing it. Tony Blair did many wrong things during his time as Prime Minister, and it seems plausible that he knew, at some level, that these actions were wrong. He is not outside of our moral community: he did not seem to have the wrong end of the stick about what morality required. Rather, he was too easily swayed by the wrong sorts of reason. He did not try hard enough. Much of his ignorance, both factual and moral, was motivated ignorance or affected ignorance. He was (and is) thus blameworthy in the ordinary way. When we blame people for their akratic acts, we take it that they have the capacities and moral knowledge that we have: they are part of our moral community in that they share the basic standards that we hold ourselves to. (Mason 2015, 12)

There are (at least) three interesting claims being made here.

1. Only people who are in our moral community are subject to ordinary blame. We might have contempt, or disdain, or disrepect for people in alien moral communities. Indeed, we might hold them subject to objective blame. Indeed, having contempt, etc for them might be a way of objectively blaming them.
2. When an ordinary person in our moral community does something wrong, they know that what they are doing is wrong, at least at some, possibly sub-conscious, level, or they are engaged in blameworthy motivated reasoning.
3. Only people who know that what they are doing is wrong, at least at some, possibly sub-conscious, level, are subject to ordinary blame, unless they are engaged in blameworthy motivated reasoning.

I'm not sure whether Mason intended to argue from 1 and 2 to 3; the text doesn't make that appear to be the central strand on her reasoning. (Cases like JoJo's are

much more important.) But as I've set this up, there is a valid argument from 1 and 2 to 3. And it's a pretty interesting argument for MIE.

Premise 2 in this argument is clearly an empirical claim. We saw a version of it in Calhoun's picture that blameless ignorance typically requires societal ignorance. People whose society generally disapproves of their moral theories have sufficient reason to temper those theories at least enough so they should not be acted on. And the general picture is continuous with what John Kenneth Galbraith was articulating when he described modern conservatism as "the search for a superior justification for selfishness" (Galbraith 1964, 16). The picture is that deep down, people who do wrong know that they are being selfish, or discriminatory, or in some way immoral, but they attempt to, or perhaps successfully motivate themselves to, justify this behaviour in moral language.

But unless the claim about moral community is taken to, by stipulation, include only people who don't have any thorough-going, unmotivated, false moral beliefs, I don't see why we should accept this claim. Here are three very broad classes of exceptions.

Hacker is an anti-abortion activist. He knows that many people think abortion is permissible, and for this reason he refrains from using violence in support of his anti-abortion crusade. But he does campaign for regulations that are used to close abortion clinics. And he hacks into the computer systems of abortion clinics to render their systems inoperative, and make it harder for them to carry out abortions. He thinks this work is morally mandatory, so he certainly thinks it is permissible. I doesn't strike me as plausible that he believes, deep down, that he is doing something wrong. We can stipulate that he knows that what he is doing reduces the autonomy of women seeking abortions. But since he regards abortion as morally equivalent to murder, he doesn't think that reducing the autonomy of would be murderers is a bad thing.

Guy is a regular American, with a regular American diet. This includes generous helpings of factory farmed meat. To the extent that he worries about this, it is just because of the health consequences. He doesn't think animals bred for food have any moral standing. (Though he does think people should be jailed for promoting dog fighting.) The interests of humans, he think, come first, and factory farming is justified because it lowers the cost of meat. It seems consistent to think that he is very badly wrong about this, and yet he doesn't have any internal sense, even deep down, that this is so.

Finally,Rush is in a hurry to get home. It's an emergency. Well, it's not an emergency really, but his child is hungry and is screaming, so it feels like one. So

Rush drives somewhat aggressively, and somewhat impolitely, and in so doing wrongs other road users. He actually does know that what he is doing is usually wrong. But he also knows that what he's doing is acceptable, and perhaps mandatory, in an emergency. (Assume that his driving, though inconvenient for other road users, would be exactly the right thing to do if he needed to race his child to hospital. So Rush really knows that there is an exception here; he's just mistaken about its scope.) And he is, like most of us, rather too willing to classify his own challenges as falling into the scope of that exception for emergency. Note that he is making a moral mistake here, not a factual one. The mistake is not about how to describe what it's like in his car. He knows that he has a hungry but not actually endangered baby who is screaming. What he's wrong about is whether this is enough to trigger an exception to the normal moral rules about carefulness and politeness on the roads. Maybe some real people who are like Rush know that there is some special pleading here. But I don't think they all do; some people make moral mistakes that are convenient without being at any level motivated.

Hacker, Guy and Rush all do something they think is permissible. And they are all wrong. And they are all in our moral community, at least as I'd understand the expression 'moral community'. And in none of those cases should their mistaken moral beliefs lessen their responsibility. That's a striking contrast with JoJo.

And this all suggests a simple explanation for the intuitions about JoJo. We intuit that there is a kind of blameworthiness, or perhaps a degree of blameworthiness, that only applies when the agent is part of our moral community. I'm not endorsing this intuition; I'm just trying to explain why we think what we do about JoJo. What I do think is that we intuit that JoJo is, in virtue of his depraved upbringing, outside our moral community in a way that his father might not be. This idea, that intuitions about blameworthiness track membership in a moral community rather than moral ignorance, seems like a better explanation of our intuitions about JoJo.

This view is similar to what Miranda Fricker (2010, 152) calls the 'relativism of blame'. She holds that

> Blame is inappropriate if the relevant action or omission is owing to a structurally caused inability to form the requisite moral thought. (Fricker 2010, 167)

I don't think this can be quite right, because moral thoughts are never requisite for an action. Fricker talks about a schoolmaster a few decades back who canes

students, but could not be expected to realise that this common practice was immoral. Well, maybe he couldn't, but he doesn't have to in order to not beat students. He just has to not beat them. Huck Finn didn't have to realise it was wrong to turn in fugitive slaves in order to not turn in Jim; he just had to not turn Jim in. But set this point aside, and assume the issue is not 'requisite' moral thoughts, but simply corresponding ones.

And I don't think we should necessarily insist that we are dealing with structurally caused thoughts. A young child with a weakly developed sense of morality may be excused for their wrong actions. And that's because they are not fully part of our moral community. But it's a stretch to call this a structurally caused inability to form moral thoughts. Better to just say that the mid-Century schoolmaster, the ancient slaveholder and the young child are not full members of our moral community, and that's why blame is inappropriate.

(Joseph Shin pointed out in a seminar at Michigan that it is an attractive feature of MIE that is offers an explanation of why children are typically exempt from moral blame. He's right about this, and this strikes me as a point that future proponents and opponents of MIE should engage with.)

But all that said, there is something plausible about Fricker's relativism of blame. To blame someone is to stand in a relationship to them that is most natural in the context of some pre-existing relationship, or at least some pre-existing commonality. People who are outside our moral community are not so much excused for their wrong actions as exempt from blame.

Moral communities are informal entities, quite unlike countries. Membership of a common moral community could be a matter of degree. So we could make a gradational version of Fricker's relativism of blame. A wrong-doer is only blameworthy to the extent they are a member of our moral community. As a corrollorary they are only fully blameworthy if they are fully a member of the community.

This approach agrees with intuitions about cases. JoJo's upbringing is so strange that he's outside our community in many respects. So he's exempt from certain kinds of blame. Hacker, Guy and Rush are in our moral community, although they make moral mistakes. Those mistakes don't exempt them from blame. On this way of thinking, the argument from cases starts with plausible premises, but overgeneralises. What is relevant is not that JoJo thinks the token acts he performs as a vicious dictator are permissible, but the broader moral system he is a part of. MIE goes wrong by focussing on local beliefs of the wrong-doer; we should be looking at structural features of their belief system.

I've said this is a plausible explanation of why we have certain intuitions, but that's a long way from saying those intuitions are true. I don't have to take a stand on that question, and I don't know what the right thing to say is. My best guess is that figuring out what to say about JoJo requires settling some very big picture questions about the role of blame in a moral theory. One possible consequence of the ideas I've just been sketching is that we shouldn't *blame* JoJo, but we should have some other negative person-level evaluation of him. That's to say, we should still in some good sense hold him responsible for his actions - though maybe not in the way that we hold people responsible when we blame them. It is not uncommon to see philosophers identify moral responsibility with susceptibility to praise and blame, and if the picture I've been building to here is right, that identification must be wrong. We could treat JoJo as responsible by, for example, being angry at him, or having contempt for him, even if we don't blame him, or think blame would be the right kind of attitude to hold. While I'm not going to settle any questions that big, I will end with some other points about blame that help make sense of the views I've defended in this chapter.

5.12. Two Approaches to Blame

Neil Levy (2005) helpfully distinguishes two approaches to responsibility.

> There are accounts that hold that an agent is responsible for something (an act, omission, attitude, and so on) just in case that agent has – directly or indirectly – chosen that thing, and there are accounts that hold that an agent is responsible for something just in case that thing is appropriately attributable to her. ... Call these accounts volitionist and attributionist accounts of moral responsibility. (Levy 2005, 2)

The view of responsibility I'm taking here is very much in the spirit of the views that Levy calls attributionist. In particular, I've relied heavily on the idea that someone can be responsible for not reconsidering their moral views. Yet we rarely choose to deliberate. Indeed, the notion of choosing to deliberate is of dubious coherence. Once we are thinking about whether to look more closely at, say, our belief that p, we already are to some extent deliberating about p. So some things that we are responsible for, namely failures to deliberate, are not choices. And that's contrary to the view that Levy calls volitionist.

The argument of the last paragraph is not novel; it draws heavily on the arguments for attributionism by Angela A. M. Smith (2005). It is sometimes thought that agents are never really responsible for certain failures, such as failures to deliberate. What they are responsible for are the actions that produce a disposition to deliberate or not in the appropriate circumstances. This seems rather implausible to me, for reasons set out by Manuel Vargas (2005). But perhaps this kind of consideration can be used to produce a form of normative externalism that is compatible with volitionism. After all, normative externalism doesn't require rejecting volitionism, at least as Levy has defined it here. Consider again the ancient slaveowner in the example that Rosen and Guerrero discuss. The slaveowner does choose to own slaves, and it is the slaveowning for which he is blameworthy. He doesn't choose to do the wrong thing as such. But it is a very stringent condition on responsibility that one choose to do the wrong thing as such. Someone could be a volitionist and a normative externalist provided they reject that stringent condition.

Levy argues that one problem for the attributionist view is that it can't distinguish between the wrong and the blameworthy. As we saw, this idea is also behind one of Calhoun's arguments. She thinks that some arguments that men are blameworthy for their part in maintaining an oppressive society turn on conflating wrong and blameworthy acts. And intuitively this is a conflation between two distinct concepts. As Levy notes, attributionists can find some difficulty in making sense of the distinction. Indeed, he quotes two prominent attributionists, Robert Adams (1985) and Gary Watson (1996) explicitly saying that thinking something is wrong is, to some extent, blaming the wrongdoer. To conclude this discussion of responsibility, I want to note that there is a possible view that holds that blameworthiness and responsibility are conceptually distinct, even though any wrong act is blameworthy. This view is particularly congenial to the normative externalist.

On the view I have in mind, wrongfulness and blameworthiness differ in three respects.

1. They have different targets. It is, in the first instance, actions that are wrong, but agents who are blameworthy.
2. They differ with respect to time. As noted above, an action can become less blameworthy over time, but does not become less wrong.
3. They frequently differ with respect to degree.

The last point is true because there are such things as partial excuses. Indeed, it is arguable that all excuses are partial excuses. By a partial excuse, I mean some-

thing that reduces an agent's blameworthiness, without fully absolving the agent. It is helpful to think of a familiar analogy from the criminal law. Sometimes, special circumstances can provide an agent with a defence; the circumstances mean that the agent was not guilty of any crime even though they fulfilled all the elements of the crime. (Defence of another is often thought of this way.) In other cases, circumstances mitigate an agent's guilt. They don't provide a reason for finding the agent not guilty, but they provide a reason for imposing a lesser punishment. In this context, a full excuse would be something that meant there was no punishment at all that was appropriate, but which did not provide a reason for finding the person not guilty. This is a strange combination. Indeed, it may be incoherent. The finding of guilt is itself a punishment. The same thing is true in the moral case. Excuses typically mitigate responsibility. But things that absolve an agent from responsibility are usually defences, which imply the agent didn't do anything wrong. Holding that the agent has no defence for what they did, but they are fully excused, is an unstable position.

The reasoning of the last paragraph suggests that the following principle is plausible:

- If S's action *X* at *t* is wrong, then S is to some extent blameworthy at *t* for *X*.

This principle does not imply that S's action is blameworthy, only that S herself is. And the principle does not imply anything about how blameworthy S is at later times. And it does not imply that S is blameworthy in strict proportion to the wrongness of her action. Indeed, none of these three claims is plausibly true. So here we have three important conceptual distinctions between wrongfulness and blameworthiness. But the principle does amount to a kind of attributionism, one that is very friendly to normative externalism. So the normative externalist, and the attributionist, need not be guilty of any conceptual confusion.

Here is another way to defend the principle. It is sufficient to count as blaming someone for an action that you in some way harm them, or sanction them, for performing the action, on non-consequentialist grounds. To believe that someone has done something wrong is to harm them. To improperly believe someone has done something wrong is indeed to wrong them (Basu and Schroeder forthcoming). It's not a wrong if the belief is well-grounded, but it is still a harm. And, like most beliefs, it isn't held on consequentialist grounds. So to believe that someone has done something is wrong is already to blame them - at least a little. When philosophers say that some wrong actions are not blameworthy, I

think it would be better to say that no further blame, beyond believing the person performed the wrong actions, is fitting. In cases like Inka's, or Gloucester's, that might be the right thing to say.

6. Double Standards

This chapter wraps up three loose ends. First, I discuss whether hypocrisy is a vice. I'm going to argue that it isn't, and say why this matters to the broader issue of normative externalism. Second, I'm going to say why I haven't relied on arguments about inter-theoretic value comparison to argue for normative externalism. Roughly, I think those arguments over-generalise, so it is a mistake to use them here. And finally, I'll say something about which aspects of normative externalism are central to the view, and which are peripheral. This will serve as a summing up of this part of the book, and help bring into focus how the different parts fit together.

6.1. Hypocrites

Janus has the following odd set of views. He has basically correct views about the physiology of animals that are used for livestock. On the basis of these views, and some straightforward philosophical reflection, he has concluded that meat eating is almost certainly impermissible. Given the philosophical evidence available to him, this isn't a particularly irrational view to have, but it is false in the world he is in. And this is good luck for Janus, since he eats meat at every opportunity.

Here's a natural objection to the simple externalism I have so far defended.

1. Janus is, in some way or other, criticisable.
2. According to simple forms of normative externalism, he is not criticisable, since what he does is not wrong.
3. So simple forms of normative externalism are wrong.

I'm going to primarily push back against premise 1 here. While there is some intuitive pull to the idea that Janus is criticisable, I will argue that intuition can easily be explained away. But first, I'll start with a couple of clarifications of the case.

6.1.1. Why hypocrisy?

Some readers may be wondering why I'm talking about Janus's hypocrisy, rather than his akrasia. After all, 'akrasia' is the standard philosophers' term for a person who acts against their better judgment. And hypocrisy is attached to doing other than what one says is right, as much as doing what one doesn't think is right.

On that last point, I think hypocrisy applies to more people than just the character who says one thing and does another. It seems fine to me to describe Robinson Crusoe as hypocritical if he comes to a firm opinion that some action is wrong, and then goes and does it. Perhaps we can understand this case as Crusoe making a speech to himself, and then being hypocritical for acting against his (inner) speech. But that feels at best like a forced reading of the case. It is more natural to say that actions can be hypocritical even if they don't conflict with any prior speech of the agent, if they do conflict in some way with her judgments.

Still, why not describe this as akrasia? Well, for one thing the term 'akrasia' is barely a term of English. There is the English expression 'weakness of will', but this has very little to do with the phenomenon we're considering here (Holton 1999). To the extent we understand what it is to be akratic, we understand it stipulatively. But as we'll see below, there are some tricky borderline cases of hypocrisy. I would like to use our intuitive judgments to help clarify those cases. But I can hardly ask the reader to share intuitions about akrasia, since it is a notion introduced by stipulation.

6.1.2. The Hypocrite and the Rationaliser

I've known the occasional person like Janus, but in many ways he seems like a rather foreign character. A much more common character, at least in the circles I typically move in, is the person who comes up with rationalizations for their particular behaviour. (I'm drawing heavily in this subsection on what Eric Schwitzgebel (2011) says about rationalizations.) It is not obvious why one would think that that making these rationalizations is a moral improvement. Indeed, I suspect I prefer the character who faces up to their own moral failures to the one who constantly finds a spurious reason to justify their own behaviour.

I only bring this up to note that if you're like me, it might be a little hard to make firm judgments about Janus. After all, Janus is so foreign, so Other, that he doesn't attract our normal sympathies, and without those sympathies we aren't

great at moral judgments. I'm not going to lean heavily on this point, but I do think that it is a reason to suspect that we might not be the best judge of people like Janus.

6.1.3. Recklessness and Character

My preferred thing to say about Janus is that his action is not in itself criticisable, but that actions like this are revealing of a character flaw that is worrying. (This diagnosis borrows from some suggestions Hawthorne and Srinivasan (2013) make about how to analyse a parallel case in epistemology.) The character flaw is not taking the interests of others as seriously as one should, especially in comparison to one's own interests. One way to do that is adopt theories or standards that are helpful to oneself. Another is simply to ignore what one thinks are the appropriate standards in one's actions.

Now Janus is not, by hypothesis, doing anything wrong. He puts his own desire for meat above the interests of the animals who are killed to provide the meat. And by hypothesis that's fine, since his interests are sufficiently more significant. But humans are notoriously bad at balancing their own interests with the interests of others. Someone who acts so as to promote their own pleasure over the interests of others, even in cases where they have judged the others' interests to be more significant, seems very prone to selfish action. We want people to act against their own interests, when the interests of others are sufficiently strong. It is true that Janus does not violate this desideratum. But since he thinks he violates it, it is likely that he will actually violate it next time he gets a chance. It is reasonable to worry about the character of such a person, even if the particular thing they are doing is not objectionably selfish.

To support this interpretation of the case, let's compare Janus to someone whose actions against their moral judgments are not self-serving. Yori is both a parent and an academic. It's hiring season, and Yori has a bunch of job applications to read. He has read them all closely, but worries that he really should go back and look at a few a bit more closely before tomorrow's meeting. But he doesn't have time to do that and attend his child's soccer game, and he knows his presence at the game will mean a lot to his child. Yori thinks that his professional obligations are stronger than his parental obligations, so he should re-read the files. But he can't bring himself to disappoint his child in this way, so he goes to the soccer game. He doesn't get any pleasure from this. He finds the soccer deathly boring. And while he would feel guilty if he skipped the game, and this feeling would not be pleasurable, as it is he feels equally bad about the files. Now it turns out

Yori has a bad theory of duty. Given the work he has already put in, his parental duties are stronger than his professional duties, so he does the right thing. And he even does the right thing for basically the right reason, being motivated by his child's feelings. (I'm assuming here that being unable to bring oneself to disappoint a child is a perfectly acceptable way to be moved by a child's feelings. If you don't agree, you may have to change the story, but for what it's worth I think that assumption is true.)

While we might criticise Yori for his false moral theory, we should not criticise his action in any way. He does the right thing, and does it for the right reason, even if he falsely believes that this very reason is not a strong enough reason. Yori is, just like Huckleberry Finn, a case of inadvertent virtue.

Yori is like Janus in one respect; he acts against his judgment of what is best to do. And he is unlike Janus in a different respect; he does not act selfishly. The appropriate attitudes to take towards Yori and Janus are very different; the kind of negative attitude that is natural to take towards Janus is uncalled for when it comes to Yori.

And this suggests that the explanation for that negative attitude towards Janus comes from the respect in which he differs from Yori, not from the respect in which the two of them are alike. That is to say, Janus is criticisable, to the extent he is, because he acts selfishly, not because he acts against his best judgment.

Selfish action is not always wrong. Sometimes, you should put yourself first. Happily, Janus is in such a situation. So why do we criticise him? It is not because he does something wrong, but because he reveals bad character. If he does something wrong, it is something that Yori too does wrong; yet Yori does nothing wrong. Janus's actions reveal a worryingly selfish personality. Even if this very action wasn't selfish, it's a good bet that he will soon act in a way that's objectionably selfish. That's not true about Yori.

And this is all evidence that hypocrisy isn't in itself a vice. Yori is hypocritical; he thinks he be reading job applications, but instead finds himself at a children's soccer game. But this isn't something bad about his actions, or even his character. Indeed, it would have been worse to act in accord with his false views about duty.

And that is the key thing for normative externalists to say about hypocrisy. Often, the hypocrite does something bad, and that should be criticised. In many other cases, the hypocrite reveals a character flaw that will, quite probably, lead to bad actions in the near future. That too should be criticised, at least with the

aim of preventing the bad actions from happening in the near future. But sometimes the hypocrite simply does the right thing, and ignores their false moral views. That isn't bad, and isn't even a character flaw. There is nothing wrong with simply doing the right thing, even if one doesn't recognise it.

6.2. Value Comparisons

There is a prominent argument against normative internalism that I have not discussed here. This is the problem of inter-theoretic value comparisons (Sepielli 2009; Hedden 2016a). I haven't discussed it because I don't think it is as big a problem for internalism as some of my fellow externalists do. But it is an interesting problem, and thinking about how it could be solved shows some constraints on the form of a viable internalism.

Ulysses is trying to decide between two problematic forms of action. If he does action A, he will break a promise to his dear wife, Penelope, but he will also improve the welfare of hundreds of people on the island he is visiting. If he does action B, he will be able to keep the promise, but he will lose the opportunity to help the people around him. Ulysses is also torn between two moral theories. One is a welfare consequentialist theory that says he should do action A, and the other is a deontological theory that says he should do action B. Let's assume that he is reasonable in being so torn. (This is a huge simplification, but the problem doesn't change with fewer simplifications, it just becomes harder to state.) What should he do?

The externalist says that we need to know whether the consequentialist or the deontological theory is correct, and that will determine what Ulysses should do. But some internalists don't like this answer. They note, correctly, that it is really hard to work out what the right moral theory is. And they think that it shouldn't be so hard to work out what to do. So Ulysses must be able to do something with just the knowledge he has. (Or so say the internalists. I obviously disagree, but we'll set aside my disagreement for the moment.)

What options does Ulysses have? If he knew one of the moral theories was more likely than the other, perhaps he could just do the thing recommended by the more likely moral theory. But we've assumed that Ulysses knows no such thing. So perhaps the thing to do is to maximise expected moral value. To do that, we just need to know whether $x > y$, where x and y are defined as follows:

- x = the amount which action A is better than action B, according to the version of consequentialism Ulysses takes seriously.
- y = the amount which action B is better than action A, according to the deontological moral theory that Ulysses takes seriously.

The problem is, how are we going to find out whether $x > y$? We can't look to either of the moral theories that Ulysses takes seriously. They can only answer questions internal to themselves; they can't say how to compare something that's wrong by their lights to a kind of wrongness taken seriously by a rival theory. What we need is a comparison of wrongness across theories. That is, we need an inter-theoretic comparison of wrongness. Or, as it is sometimes put, we need an inter-theoretic value comparison. (It sometimes seems to me that there is an implicit consequentialism built into this way of putting the problem, but set that worry aside.)

Now there are a number of moves that have been proposed for how to get around this impasse, and a number of criticisms of each of them. What I'm interested in here is the following argument.

1. If normative internalism is true, there is a solution to the problem of inter-theoretic value comparison.
2. There is no solution to the problem of inter-theoretic value comparison.
3. So, normative internalism is false.

Premise 2 of this argument is false. I don't say that because I know the solution, or because I have an argument in favour of a particular solution. What I do have is an argument that a solution must exist. The argument turns on considerations about democracy and representation.

Saraswati is a good democratic representative. She currently faces a tricky decision between two options. One option will maximise welfare, but breach some moral principles that are often held to be important. The other option will do neither of these things. Now as it turns out, the true moral theory in Saraswati's world is a kind of pluralism that says it is morally permissible to make either choice when acting for oneself, and making this kind of choice. But Saraswati isn't acting for herself, she is a representative. And representatives have a duty to represent, at least in cases where the people want them to act in morally permissible ways. And it turns out Saraswati's constituents are torn. Half of them are committed welfarist consequentialists, the other half are deontologists. Assume further that Saraswati has not promised, either implicitly or explicitly, to make one choice rather than the other in this kind of situation.

Given all those assumptions, what Saraswati should do turns on the correct answer to the problem of inter-theoretic value comparison. What she should do depends, at least in part, on whether the welfare loss matters more to her welfarist constituents than the principle violation matters to her principled constituents. That is to say, what she should do turns on exactly the same kind of question that Ulysses faced when he was deciding whether $x > y$.

Now as an externalist, I don't think Saraswati has to in any sense solve the problem of inter-theoretic value comparison. She just has to do the right thing for the right reasons. And one can do the right thing for the right reasons without knowing they are the right reasons, or even without having any general disposition to act rightly in similar cases. But I do think that we as theorists need to solve the problem in order to say anything evaluative about Saraswati's actions. And even if we can't do that, if we believe there is a fact of the matter about whether Saraswati did the right thing, then we are committed to thinking that there is a solution to the problem of inter-theoretic value comparisons.

So premise 2 in the externalist argument above is not true. There is a way, somehow, of solving the problem of inter-theoretic value comparisons. That's not to say it will be easy. Personally, I suspect it is one of the hardest problems in all of ethics. None of the remotely viable solutions to it seem either obvious to the lay actor, or easy to implement.

The difficulty of solving the problem does not show that normative internalism is false. But this difficulty does undermine a popular motivation for internalism. If the idea behind internalism is that there should be a sense of 'should' in which ordinary people can usually tell what they should do, that can't be a sense of 'should' which is sensitive to the correct solution to the problem of inter-theoretic value comparisons. So it is incoherent to motivate internalism by saying that externalism makes it too hard to know what to do, and then develop a theory of right action that requires a solution to the problem of inter-theoretic value comparisons.

As we've seen, this isn't the only way to motivate internalism. Some theorists motivate internalism by an analogy to the wrongness of reckless action. Those theorists often need there to be a solution to the problem of inter-theoretic value comparison, but they don't need this solution to be in any way transparent. And reflection on cases like Saraswati's makes me think that the externalist must concede that such a solution must exist, and so cannot rely on its non-existence in arguing against internalists.

It's worth noting just how strong a conclusion we could draw from the inter-theoretic value comparisons argument. Assume, for reductio, that we really couldn't make sense of any kind of inter-theoretic value comparison. It would follow that there is no way to define hypocrisy in probabilistic terms. Someone could only be counted as a hypocrite if they fully believed that what they were doing was wrong. But this doesn't seem right.

Imagine that someone faces a choice about whether to betray a confidence. The betrayal would be extremely disrespectful, but they think there would be a small gain to the welfare of the world if they did so. And while they mostly think respect is central to morality, then have a non-zero credence that welfare consequentialism is the correct moral theory. They break the confidence. Are they hypocritical? I think they probably are, even if we can't give an algorithm for weighing the downside of the betrayal on the moral theory they think is probably right against the welfarist upside on the theory they give a small credence to. If we think the problem of inter-theoretic welfare comparisons is not solvable in principle, then we can't even define a notion of hypocrisy that applies in cases like this. And that would be a very strong result. I don't think the arguments that hypocrisy is no vice are nearly as strong as the arguments against more systematic forms of internalism in chapters 2–4. So I suspect the argument from inter-theoretic value comparisons proves too much. It doesn't just rule out views like *The best thing to do is maximise expected goodness*, it also rules out views like *It's at least a minor vice to not live up to your own principles.* And that feels like too much weight for the argument to bear.

6.3. The Externalist's Commitments

I'm going to finish up this part by saying a bit about what I take to be the more and less central parts of normative externalism. Like any -ism, the view not only makes many different commitments, those commitments differ greatly in strength. Setting out these commitments serves a few useful functions. It helps us see how the different parts of the view hang together. And it is good practice to say ahead of subsequent refutations what retreats would be minor setbacks, and what would amount to fleeing the field[1]. It's very tempting when a part of one's view is shown to be flawed to insist that it was only a peripheral aspect of the view to begin with. Writing down which commitments are central

[1] I'm assuming here that there will indeed be subsequent refutations, but this follows from a version of the pessimistic induction.

and which are peripheral before the flaws are made visible is a way to avoid this temptation.

The core idea is that moral norms are independent of both what one thinks the moral norms are, and what one should think the moral norms are. Here are a few ways that could fail that would threaten the periphery of the view; we'll then move to seeing what a catastrophic failure would look like.

First, there could be some one-way dependencies between moral norms and (rational) beliefs about moral norms. For instance, a view that said being true to oneself was one moral requirement among many would violate one direction of the externalist's independence constraint. It would say that believing that something is wrong is sufficient, but not necessary, to make performing the action wrong. And the view discussed in chapter five, where believing that an action is not wrong excuses it, violated the other direction. It says that believing that something is wrong is necessary, but not sufficient, for the performance to be blameworthy.

Second, there could be some dependencies that concern minor aspects of morality. The most natural versions of this possibility combine it with the one-way dependencies of the previous paragraph. Consider a view that said that hypocrisy is a minor vice. Such a view might say that it's bad to do what you think is wrong, but unless the belief is true, this is not a major vice. Or consider a view where false moral beliefs are partial excuses. These are paradigms of the 'peripheral' failures I was talking about above. What I want ultimately to argue for is that morality is about respect, welfare, rights and so on, and not about conformity to one's own principles. A view that says that morality is almost entirely about respect, welfare, rights and so on, but conformity to one's own principles has a small role too, is inconsistent with my preferred view, but the differences are minor.

The third kind of peripheral failure takes a little more setup. Consider again Descartes' view of the good person, and compare it with Kant's view. (I'm simplifying both thinkers here, but the caricatures are useful for setting out the philosophical point.) Both of them think that the good person will do what they think is right. But Descartes thinks this because he thinks that resoluteness is one of the supreme virtues. Kant, on the other hand, thinks this because he thinks that the nature of the moral law is visible to good people. So because the moral law is the way it is, the good person will both act a certain way, and have correlated beliefs about morality. For Descartes, the fact that the person beliefs that they should do X explains why it is good that they do X. For Kant, the fact that the

moral law is the way it is explains both why it is good that the person does X, and good that they believe that X is good to do. In Descartes' case, but not Kant's, the moral beliefs explain the moral status of the action.

More generally, we can distinguish amongst views that say there is a connection between morality law and what one (reasonably) believes about morality. Some such views say that (reasonable) beliefs about morality explain why actions have the moral status they do. Other views say that the moral status of the actions explain why certain beliefs are actual or, more likely, reasonable. Yet other views say that some third thing explains both the moral status of the actions and the actual or reasonable beliefs about their moral status. What I'm really committed to denying is the first of these options, where actual or reasonable beliefs about morality explain the moral status of actions.

We can put all this in terms of a checklist. Ideally, from the point of view of normative externalism, there would be no necessary connections between moral properties, on the one hand, and actual or reasonable moral beliefs on the other hand. If, however, there is such a connection, we can ask three questions about it.

1. Is the connection two-way, as opposed to moral beliefs providing merely a necessary or a sufficient condition for the moral property?
2. Is the moral feature morally central, as opposed to being, say, a minor vice or virtue?
3. Does the (actual or reasonable) moral belief explain why the moral property is instantiated, as opposed to the explanation going the other way, or some third factor explaining the connection?

The more 'yes' answers we give, the worse things are for normative externalism. The view I want to defend is that there are no necessary connections between moral belief and morality. But if there are such connections, I want to defend the view that these are one-way, or are minor, or that the moral belief does not explain the moral property. Being true to yourself is not part of morality. But if it is, it is a small part, and actions that are true to yourself aren't good because they are true to yourself.

This last possibility, the one about surprising orders of explanation, is a useful segue into epistemology. Here's another way that normative externalism could strictly speaking fail, without threatening the core commitments of the theory. There has been a pronounced 'factive turn' in recent epistemology. Many epistemologists think that our most important epistemological concepts are factive.

The most important ways for beliefs to be good are such that if a belief is good in that way, it must be true. One way to implement the 'factive turn' is to make knowledge central to epistemology. But another way, not inconsistent with the first, is to argue that other epistemological notions are factive. And that turns out to have consequences for normative externalism.

Let's say that one thought, on quite general grounds, that only true beliefs could be rational, or that only truths could be well supported by evidence. I don't think either of those things, and I'll say a little more in the next part as to why, but for now I just want the view on the table. That would imply there is a necessary connection between rational moral beliefs and morality. If one rationally believes that lying is wrong, then it must be that lying is wrong. But that's not because the rationality of the belief explains the wrongness, or that the having of the belief explains the wrongness. It's because the wrongness of the lying is a necessary precondition of rationally believing that lying is wrong.

This kind of view would not show us anything special about morality. On such a view, if one rationally believes that lying is common, then it must be that lying is common. And it doesn't threaten the central commitments of normative externalism. But it does mean there is a necessary connection between morality and rational moral belief. So it's a small defeat, but one we can absorb without too much distress. When we turn to epistemology, we'll have to pay more attention to this kind of possibility.

Part II.

Epistemology

7. Level-Crossing Principles

7.1. First-Order and Second-Order Epistemology

In the previous part I argued that morality is independent of both what one thinks about morality, and what one should think about morality. In this part I want to argue the same thing for epistemology. But we have to be a bit careful setting up the independence thesis. Informally, a thesis of the previous part was that morality and epistemology are distinct existences. Arguing 'the same thing' for epistemology would amount to arguing that epistemology and epistemology are distinct existences. That doesn't sound particularly plausible. So to state my intended conclusion a bit more carefully, and a bit more plausibly, we need one bit of terminology.

Say that a claim that either describes or evaluates a particular belief of a person is first-order when that very belief is not itself a description or evaluation of a particular belief. And say that a claim that either describes or evaluates a particular belief of a person is second-order when the belief in question is a description or evaluation of another belief. So here are some examples of first-order claims.

- Baba believes that his keys are missing.
- Baba should believe that his keys are missing.

And here are some examples of second-order claims.

- Baba believes that he believes that his keys are missing.
- Baba should believe that he believes that his keys are missing.
- Baba believes that he should believe that his keys are missing.
- Baba should believe that he should believe that his keys are missing.

And we can replace 'should' in any of these claims with any other kind of epistemic norm. So here are some more first order claims.

- Baba's evidence supports the belief that his keys are missing.
- Baba's belief that his keys are missing is justified.

- Baba rationally believes that his keys are missing.

And here are a sample of some more second-order claims.

- Baba should believe that he rationally believes his keys are missing.
- Baba's evidence supports the belief that his belief that his keys are missing is justified.
- Baba's belief that he believes that his keys are missing is justified.

The core thesis of this part of the book, the core thesis of normative externalism in epistemology, is that first-order and second-order claims are independent. There are no true level-crossing principles, describing necessary connections between first-order and second-order claims.

Just like in part one, there are fall-back positions I will adopt if this strong claim (no necessary connections at all) turns out to be false. If there are necessary connections, they are one-way, or they are about less central concepts in epistemology, or the explanation of the claim does not go from the second-order claim to the first-order claim. But I'd rather not retreat even to there, and instead to argue that there are no true level-crossing principles at all.

I'm interested in level-crossing principles for a few reasons. For one thing, I find them intrinsically interesting. For another, they have consequences for a bunch of epistemological debates. I'm going to discuss at the end of the book the consequences they have for disputes about how to best respond to peer disagreement. But they also matter for a bunch of other debates. And they matter because to the extent they are true, they push us towards a certain kind of coherentism, and away from a certain kind of foundationalism. Just which kind will depend on just which level-crossing principles are true. But the general idea is that if rationality requires conformity to one's own beliefs about the rational, then rationality is more of a coherence concept than we might have thought it was.

7.2. Change Evidentialism

It isn't just the principles that push away from foundationalism. The examples that are used to motivate level-crossing principles are also taken to mitigate against a fairly weak form of foundationalist evidentialism that I'll call Change Evidentialism.

Change Evidentialism A person with a rational attitude towards p is under no rational obligation to change that attitude unless their evidence for or against p changes.

I think Change Evidentialism is true. Indeed, I think a much stronger form of foundationalist evidentialism, one that says the rational status of a mental state supervenes on the evidence it is based on, is true. It is far beyond the scope of this book to defend the stronger claim. I will, in effect, be defending evidentialism against a class of attacks, but that defence will not be the focus.

Change Evidentialism is related to theses level-crossing principles because some cases that motivate the principles also appear to undermine Change Evidentialism. Here is one such case, due to David Christensen.

> I'm a medical resident who diagnoses patients and prescribed appropriate treatment. After diagnosing a particular patient's condition and prescribing certain medications, I'm informed by a nurse that I've been awake for 36 hours. Knowing what I do about people's propensities to make cognitive errors when sleep-deprived (or perhaps even knowing my own poor diagnostic track-record under such circumstances), I reduce my confidence in my diagnosis and prescription, pending a careful recheck of my thinking. (Christensen 2010a, 186).

We might naturally reason about the case as follows. (Note this isn't Christensen's own considered take on the case.) When the resident learns he has been awake 36 hours, he does not get evidence against the diagnosis. That a particular resident has been awake awhile seems evidentially irrelevant to whether a particular patient has, let's say, dengue fever. But it is rational, indeed it is rationally required, for the resident to change his attitude towards the diagnosis on learning how long he's been awake. That's a counterexample to Change Evidentialism. And the explanation for why rationality requires a change is, we might conjecture, that principle 1 is true. The resident does have evidence excellent that he's making irrational diagnoses. So he can't rationally believe that he rationally believes the diagnosis. So, by the contrapositive of 1, he can't rationally believe the diagnosis.

I'm going to argue that the previous paragraph is all false. What the resident should do depends a lot on the details of the case. On some ways of filling in the case, the resident's evidence changes substantially, so Change Evidentialism is

consistent with the resident rationally changing their view. Indeed, the explanation of the change of view in terms of change of evidence is preference to the explanation in terms of a level crossing principle like 1. That's because in other versions of the case, where the resident's evidence does not change, the belief in the diagnosis should not change either.

So I'm going to argue that cases like Christensen's resident not only fail to challenge Change Evidentialism, they end up supporting it. And because the cases support it, they don't support 1–4. There are also direct counterexamples to 1–4. The example in the next chapter of Roshni is one such counterexample.

7.3. Motivations for Level-Crossing

The rest of this book will be devoted to investigating three recent motivations for level-crossing principles. The first concerns higher-order evidence, the second akrasia, and the third peer disagreement.

7.3.1. Higher-Order Evidence

As well as evidence that bears on a question, agents can have evidence that bears on the rationality of their verdicts about the question. Christensen's example involving the medical resident is one such case. Elsewhere, Christensen has provided several other examples along similar lines, e.g., (Christensen 2007a, 8), (Christensen 2010b, 126) and (Christensen 2011, 5–6). Similar examples have also been proposed by Adam Elga (2008), Thomas Kelly (2010, 140), Joshua Schechter (2013, 443–44) and Sophie Horowitz (2014, 719). The examples suggest something like the following argument against Change Evidentialism.

1. It is irrational for the resident in this case to stick with the original prescription without making some kind of cross-check.
2. The best explanation of why it is irrational to stick with the original prescription is that it is irrational to stick with the original diagnosis, i.e., the original belief.
3. The information the nurse provides is not evidence one way or the other about whether the patient has the disease originally diagnosed.
4. It was rational, before the information the nurse provides about how long the resident has been awake, to believe in the original diagnosis.

5. So this is a case where the rationality of a belief changes without any change in the evidence.

The last line follows from what came before, so the issue is whether the first four claims are true. I'm going to raise doubts about every one of those steps. But this is, I think, the most pressing challenge to Change Evidentialism.

7.3.2. Akrasia

Assume, for reductio, that the level-crossing principles are false. And assume that in any field, it is possible to have evidence that supports being extremely confident in something that is, as a matter of fact, false. Then there should be cases where one's evidence strongly supports *p*, but one's evidence also strongly supports the falsehood that one has very poor evidence for *p*. If one follows the evidence where it leads, one should be very confident in is the conjunction *p, and I have very weak evidence for p*. Assuming one believes (correctly!) that it is rational to follow the evidence where it leads, one should believe the conjunction: *p, and it is irrational for me to be confident in p*. But it is absurd to think that one can rationally be confident in either of these conjunctions; they are instances of epistemic akrasia, and akrasia is paradigmatically irrational.

I'm going to come back to this argument in chapter 10. The main response will be that the apparent absurdity is really not that absurd. Indeed, the intuition that it is absurd can be shown to be highly unreliable; it supports the 'absurdity' of many things that are plainly true. For now, note the connection between intuitions about akrasia and intuitions about Christensen's resident case. If the resident follows the evidence where it leads, he'll believe that the diagnosis is correct, and this belief is irrational. It looks like Evidentialism, and perhaps just Change Evidentialism, implies that the resident should be akratic. Unlike many philosophers, I won't take this to be a decisive objection to Change Evidentialism.

7.3.3. Disagreement

It seems possible for people who are known to have equally good track records, and who in some sense have the same evidence, to come to different conclusions. When they do, there is something intuitively plausible about each moving their beliefs in the direction of the other. Here is one such case.

> Ankita and Bojan have known each other for a long time, and know each other to be equally reliable, and equally reasonable, when it comes to arithmetic problems about as complex as multiplying two two-digit numbers. For some practical purpose they need to know what 22 times 18 is. They each do the multiplication quickly in their head. Ankita announces that she got 396, while Bojan announces that he got 386. (Compare a similar case in Christensen (2007b, 193).)

Again, we can use the case to construct an argument against Change Evidentialism, as follows.

1. Ankita's original evidence provides her strong reason to believe that 22 times 18 is 396.
2. Bojan's announcement is no evidence against the claim that 22 times 18 is 396.
3. Yet, on hearing Bojan's announcement, and respecting the fact that the two of them have equally good track records, Ankita should be unsure which of them is right, and which wrong, on this occasion.
4. Since Ankita knows what each of them announced, the only way she can consistently be unsure which of them is right is to be unsure whether 22 times 18 is 396.
5. So although Bojan's announcement does not change her evidence that bears on whether 22 times 18 is 396, it does change whether it is rational for her to fully believe that this is true.

Again, this looks like a reasonably intuitive argument against Change Evidentialism. And again, I'm going to raise doubts about every premise. The focus of chapter 12 will be the picture of disagreement behind premise 3. This is the view that has come to be called *conciliationism*. But what I say about evidence over the next few chapters will also raise concerns about the first two premises of the argument.

7.4. The Plan for the Rest of the Book

In what follows, the even-numbered chapters will deal with the three big arguments for level-crossing principles, and against Change Evidentialism, that I just

discussed. In chapter 8, I'll discuss higher-order evidence; in chapter 10, I'll discuss akrasia principles, and in chapter 12, I'll discuss disagreement. In between I'll address two big issues that arise out of those discussions.

In chapter 9, I'll talk about what it means for some reasoning to be problematically circular. This turns out to matter to our purposes because of a potential bit of circular reasoning that is, according to my view, perfectly acceptable. In particular, in some cases where there is reason to believe the agent is incapable of correct reasoning, I think it is possible for the agent to simply do some correct reasoning, notice that it is correct, and infer that they are, after all, capable. This can feel worryingly circular. But it turns out to be incredibly hard to find an anti-circularity principle that is both true, and violated by this reasoning.

In chapter 11, I discuss how level-crossing principles lead to nasty regresses. More precisely, I argue that level-crossing principles are only motivated if one accepts a particular assumption concerning evidential screening. And that assumption, I argue, leads to nasty regresses. The regress arguments here are similar to the regress arguments in chapter 2 against very strong level-crossing principles in ethics.

And in chapter 13, I briefly summarise the lessons of both the epistemology part, and of the book as a whole.

But before we get to all that, I need to do a little ground-clearing. The rest of this chapter contains two fairly self-contained sections on things I wanted to get out of the way before defending Change Evidentialism against level-crossing principles. The next section concerns the relationship between state-level evaluations, like the rationality of a belief, and agent-level evaluations, like the wisdom of a believer. And then I argue, against most orthodox wisdom in epistemology, that we acquire evidence while doing mathematical investigation.

These two sections are helpful for understanding the rest of the book. But they are not essential. And someone who is impatient to get on to higher-order evidence, akrasia or disagreement could skip ahead to any one of those chapters.

7.5. Evidence, Rationality and Wisdom

Change Evidentialism is a claim about the rationality of beliefs and other doxastic attitudes. The level-crossing principles I reject are principles about evidence, and about rationality. The focus here, as you may have gathered, is on rationality

and on evidence. There are other concepts in the area that I don't have as much to say about, and which may not be systematically related to those concepts.

I don't want to assume that a belief is rational if and only if it is justified. It might be that only true beliefs are justified (Littlejohn 2012), but it is very unlikely that only true beliefs are rational.[1] In any case, there is something a little artificial about talking about justified beliefs. In everyday English, it is typically actions that are justified or not. The justification of belief seems a somewhat derivative notion. So I'll stick to rationality.

I'm also going to set aside, for the most part, a discussion of wisdom. Just as in the discussion of ethics, it is very important to keep evaluations of agents apart from evaluations of acts or states. It is attitudes or states that are in the first instance rational or irrational. We can talk about rational or irrational agents, but such notions are derivative. Rational agents are those generally disposed to have rational attitudes, and to be in rational states. Wisdom, on the other hand, is in the first instance a property of agents. Again, we can generalise the term to attitudes or states. A wise decision, for instance, is one that a wise person would make. But the wisdom of agents is explanatorily and analytically prior to the wisdom of their acts, judgments, decisions and attitudes.

I think that everything I said in the last paragraph is true if we use 'wise' and 'rational' and their cognates with their ordinary meaning. But I'm not committed to that, and it doesn't matter if I'm wrong. You can read me as stipulating that 'rational' is to be used as a term that in the first instance applies to states, and 'wise' is to be used as a term that in the first instance applies to agents, and little will be lost.

Change Evidentialism is not a claim about wise agents, it is a claim about the rationality of various beliefs and belief transitions. Perhaps a wise agent is one who always has rational attitudes. If so, then Change Evidentialism will have some implications for what wise agents are like. But it is far from obvious that wisdom and rationality are this tightly linked. Indeed, at the end of chapter 11, I'll come back to a reason to question the connection. For all I've said, it may well

[1] That false beliefs can be rational seems more plausible to me than the premises of any argument I could give for it. But here is one independent way to make the case. Arbitrarily high credences in false propositions can be rational. Indeed, false propositions can have arbitrarily high objective chances, consistent with those chances being known. In such cases the only rational credence matches the chance. The best theories of the relationship between credence and chance do not require credence 1 for belief simpliciter (Weatherson 2014a). And if a high credence constitutes a belief, and the credence is rational, the belief is rational. So some false beliefs can be rational.

be wise to change one's beliefs in some situations where one's evidence does not change. That is consistent with Change Evidentialism, provided we understand those situations as being ones where it is unwise to have rational attitudes.

I am leaning heavily her on work on the connection between rationality and wisdom by Maria (Lasonen-Aarnio 2010b, 2014a). I agree with almost everything she says about the connection. The biggest difference between us is terminological. She uses 'reasonable' and 'reasonableness' where I use 'wise' and 'wisdom'. In my idiolect, I find it too easy to confuse 'rational' and 'reasonable'. So I'm using a different term, and one that, to me at least, more strongly suggests a focus on agents not states. But this is a small point, and everything I say about the distinction draws heavily on Lasonen-Aarnio's work.

7.6. Evidence, Thought and Mathematics

The picture of evidence behind the version of evidentialism that I'm presenting here differs from a natural picture that many epistemologists have. In particular, I draw the line between acquiring evidence and processing evidence at a very different place than many others do. I'm going to motivate this re-drawing by working through some examples involving mathematics. Much of what I say about these examples follows closely the arguments that Paul Boghossian (2003) made against simple forms of reliabilism and internalism about logic and mathematics. But these arguments of Boghossian's are worth rehearsing, because their significance for recent epistemological debates has not always been appreciated.

A young mathematics student, Tamati, starts thinking about primes. He notices the gaps between primes get larger, and starts to wonder whether there is a largest prime. He is struck by a sudden strong conviction that there is no largest prime, and so forms the belief that there is no largest prime. Now Tamati is not usually prone to forming beliefs on the basis of spontaneous convinctions like this. Apart from this time, he only does this for very simpile arithmetic claims, like that seven plus five is twelve. But nor is he a mathematical savant. He couldn't produce any reason for the claim that there is no largest prime. He hasn't seen, even implicitly, anything like the argument that if n is the largest prime, then $n! + 1$ would be both prime and not prime. It's just an immediate convinction for him.

Tamati does not know that there is no largest prime. This fact, assuming it is a fact, needs explaining. The evidentialist has a natural explanation. In a normal case, when someone comes to learn by proof that there is no largest prime, there are two extra facts they learn. The first is that if n is the largest prime, then $n!+1$ is prime; the second is that if n is the largest prime, then $n! + 1$ is not prime. These in turn aren't immediately obvious; to be known they must be figured out on the basis of other things. Those extra pieces of knowledge are extra evidence[2] It could be that the extra evidence would just be the premises that he would use, not the conclusions he draws from them. Or it could be that we need to give up the idea that evidence is non-inferential. I don't have a worked. It is with that evidence that a normal student can come to know, by proof, that there is no largest prime. Alternatively, the student may learn that some teacher, or some book, says that there is no largest prime, and that teacher, or book, is reliable. Those things are the extra evidence. That case is clearly different to Tamati's, because it relies on engagement with the outside world. But even the student who thinks through the case themselves acquires evidence, namely the above facts about the relationship between n and $n! + 1$.

So the evidentialist has a nice explanation of what is going on in Tamati's case. Other explanations look less promising.

We could try to explain Tamati's case in strictly reliabilist terms. But note that Tamati's convictions are perfectly reliable. The method 'trust my convinctions' gets him arithmetic knowledge every day, and the true belief that there is no largest prime. In no case does it go wrong. So the reliabilist has no explanation of why this use of Tamati's convictions does not yield knowledge. The reliabilist could try to argue that methods have to be individuated more finely than this; it is different to trust one's convictions about simple matters as compared to more complex matters. But this assumes we have some grasp on the idea that saying there is no largest prime is a complex matter. It isn't clear why this should be so. It isn't hard to state the proposition that there is no largest prime. It is a little hard to prove it. The evidentialist has an explanation of why how hard it is to prove the theorem matters to whether Tamati can spontaneously know it. But it seems very hard to motivate the idea that proof complexity should define the

[2]Is this consistent with my earlier note that we would, for the sake of discussion, identify evidence with non-inferential knowledge? It isn't obvious that it is, and it is more than a little tricky to say just what evidence Tamati would gain if he worked through the problem carefully. It would work to defend normative externalism if we identified evidence with all knowledge, as Williamson suggests, but I would rather not make that identification on other grounds. I hope to return to the question of just how we should conceptualise evidence, both in mathematical and empirical investigations, in subsequent work.

relevant reference class. It seems to use the very thing we were trying to give a reliabilist explanation of. In any case, even if we restrict the reference class to things that are hard to prove, Tamati's convictions are still reliable. He sensibly declines to form beliefs about most things in this class, while forming one true belief. So he's got a perfect success rate, so is reliable!

Alternatively, we could say that Tamati doesn't "appreciate" the evidence for the absence of a largest prime. (The idea that appreciating the evidence is important to mathematical knowledge comes from Richard Fumerton (2010), though he doesn't use it for quite this purpose, and shouldn't be thought responsible for the view I'm about to criticise.) The thought would be that Tamati has some evidence about primes, but doesn't stand in the special relationship to it needed to ground knowledge. This is obviously an anti-evidentialist position, since it says that rationality depends not just on what evidence one has, but on some further relationship that one may or may not stand in to evidence.

But depending on how we understand 'appreciate', the view will be too strong or too weak. If appreciation means understanding how and why the evidence supports the conclusion, and appreciation is required for knowledge, then very few people will know very much. Before they take a logic class, introductory students can come to know Ga by inferring it from Fa and $\forall x(Fx \rightarrow Gx)$. But they don't need to know how or why their evidence supports Ga. Indeed, they can be radically mistaken about the nature of logical implication, as many students are, and still know Ga on that basis. On the other hand, if appreciation means having a true belief that the evidence supports the conclusion, it won't rule out Tamati knowing that there is no largest prime. We can assume that Tamati is sophisticated enough to know that mathematical truths are entailed by any proposition. So if he believes there is no largest prime, he can immediately (and correctly) infer that the fact that his coffee has gone cold entails there is no largest prime. But that isn't enough for him to know there is no largest prime, not even if he knows that his coffee has gone cold. If we insist that appreciation means knowing that the evidence supports the conclusion, then we are back where we started, needing to explain why Tamati doesn't know that there is no largest prime.

So the best explanation of Tamati's ignorance is that he lacks sufficient evidence to know that there is no largest prime. If he worked through the problem slowly, he would acquire evidence for that conclusion. And that's the general case. Thinking through a mathematical problem involves acquiring mathematical evidence. Similarly, when one has to do some mathematical reasoning to get from empirical data to empirical conclusion, that reasoning doesn't just involve

processing the empirical evidence, it involves acquiring new, mathematical evidence.

This way of thinking about mathematics is hardly radical. It is a commonplace in mathematical discussions that one can get evidence for or against mathematical propositions. Philosophers too often think that evidence that entails a conclusion is maximally strong evidence. This assumption is even encoded into probabilistic models of evidential support. But it isn't true. Facts about Andrew Wiles's diet are terrible evidence that Fermat's Last Theorem is true, although they entail it. Fact about what he wrote in his notebooks, on the other hand, are excellent evidence that it is true. Thinking that entailing reasons are maximally strong reasons is just another way to confuse inference with implication (G. Harman 1986).

This attitude, of thinking that entailing reasons are maximally strong reasons, goes along with another bad attitude that it is easy to adopt. That is the attitude that when *p* is a mathematical proposition, our evidence supports either a maximally strong belief in *p* or a maximally strong belief in $\neg p$. There are numerous counterexamples to this view. Sanjoy Mahajan (2010) describes heuristics that can be used to quickly refute various mathematical hypotheses. The heuristics involve, for example, checking whether the 'dimensions' of a proposed identity are correct, and checking limit cases. So consider the hypothesis that the area of an ellipse is πab, where a is the distance from the centre of the nearest point on the ellipse, and b is the distance from the centre to the furthest point. After going through a number of other proposals and showing how they can be quickly refued, Mahajan says this about the proposal that the area is πab,

> This candidate passes all three tests ... With every test that a candidate passes, confidence in it increases. So you can be confident in this candidate. And indeed it is correct. (Mahajan 2010, 21)

It might be worried that the position I'm adopting here, that we often need evidence of a connection between premises and conclusion in order to reasonably infer the conclusion from the premises, even when the premises entail the conclusion, risks running into the regresses described by Lewis Carroll (1895). It certainly would be bad if my view implies that to infer q from p and $p \rightarrow q$, and agent needed to know $(p \wedge (p \rightarrow q)) \rightarrow q$. That way lies regress, and perhaps madness. But that's not what my view implies. The claim is just that for non-obvious entailments, the agent needs extra knowledge to infer from premises to conclusions. It is consistent with this to say that immediate entailments, like

modus ponens, can justify immediate inferences. And that's enough to stop the regress.

The idea that we accumulate evidence when working through philosophical or mathematical puzzles will matter quite a bit for debates about disagreement. It is agreed on all sides that when the parties to a disagreement do not have the same evidence, then the existence of the disagreement is a reason for each to move their attitudes. (Assuming, of course, that the other person is not irrational, or known to be bad at processing this kind of evidence.) If we allow that there is philosophical evidence, then it will be incredibly rare that each party to a debate has the same evidence. It will be vanishingly rare that each party knows that each party has the same evidence. This means that any case where the parties know about the evidence the other parties have will be a fair way removed from the kind of real-world case where we have reliable intuitions. It also means that in practice learning about the existence of a people who disagree with you is often evidence that there is evidence against your view that they have and you lack.

The main claim I'll need in what follows is that thinking through a case sometimes gives you evidence. But it's independently interesting to think how far this extends; to think about how much reasoning is a form of evidence acquisition. And examples with the same structure as Tamati's can be used to motivate the thought that very often reasoning involves evidence acquisition.

A, B and C are trying to figure out how many socks are in the drawer. They each know there are seven green socks, and five blue socks, and that that's all the socks, and that no sock is both green and blue. From this information, they all infer, and come to know, that there are seven plus five socks in the drawer. A is an adult with statistically normal arithmetic skills, so she quickly infers that there are twelve socks in the drawer. B is a three year old child, who is completely unreliable at arithmetic. She guesses that there are twelve. At this stage, we can say that A knows there are twelve socks in the drawer, and B does not know it. But C, who is four years old, is a more subtle case. She says to herself, "I think it's twelve, but I better check." That's a good reaction; like B she isn't so reliable that she can know without checking. So she uses a method for doing addition; she starts counting from seven, putting one finger up at each count. So she says "eight" and raises her thumb, "nine" and raises her index finger, and so on through saying "twelve" and raising her little finger. She looks at her hand, sees that she has five fingers raised, and concludes the answer is twelve.

At the end of this process, but not before, she knows that there are twelve socks

in the drawer. Indeed, it is only at the end of the process that she is in a position to know that there are twelve socks in the drawer. It is because she has come to know that seven plus five is twelve that she has sufficient evidence to know there are twelve socks in the drawer. Previously, this was not part of her evidence, now it is, and now she can know there are twelve socks in the drawer. That suggests it is because A knows that seven plus five is twelve that she can know there are twelve socks in the drawer to. She might not have consciously said to herself that seven plus five is twelve, but if she didn't have that as part of her evidence, she wouldn't have been in a position to know that there were twelve socks in the drawer. C also knows this, because she acquired this evidence. Indeed, she acquired it a posteriori; it in part relied on seeing that she had five fingers raised.

So even reasoning that relies on simple arithmetical identities relies on those identities being in evidence. In these cases, the only rule of implication that really seems to do double duty as a rule of inference is the transitivity of identity. An agent who knows that x equals seven plus five, and knows that seven plus five equals twelve, is in a position to infer that x equals twelve. They don't, it seems, need to know that identity is a transitive relationship. Whether we grant C knowledge that there are twelve socks in the drawer does not, it seems, depend on whether we grant her knowledge of the fully general principle that identity is transitive.

What's special about that last step is that the general principle that might be relevant is considerably more complicated to state, and to believe, than the general principle in arithmetic cases. It is easier to know that seven plus five is twelve than it is to know exactly what rule about identity that A, B and C need to use to figure out how many socks there are in the drawer. It is harder to know the general principle of disjunctive syllogism than it is to use it on an occasion. So it might be that there are more kinds of simple inferential steps that track simple implicative rules than there are kinds of simple inferential steps that track simple arithmetic identities. For all I've said here, it might be that all arithmetic inferences just involve the transitivity of identity, plus knowledge of a lot of arithmetic facts. It isn't so plausible that all logical inferences just involve one rule, such as modus ponens, plus a lot of logical facts.

8. Higher-Order Evidence

8.1. Varieties of Higher-Order Examples

Higher-order evidence is evidence about one's own evidence, or reliability, or rationality. Several examples have been proposed which are often taken to show that rationality requires adjusting one's confidence in certain propositions to higher-order evidence. And the best explanation of that phenomena may well be that some level-crossing principle or other is true. Since it's my task to argue against level-crossing principles, I need to say something about these examples.

The examples that have been proposed thus far all have a similar structure. The hero starts out with a firm belief, and the belief would licence a decisive action. Something happens that would, in normal cases, cause a person to question both the belief and the wisdom of taking decisive action. The suggested explanation is that a level-crossing principle is true, and explains the normal person's hesitation. But the structure of the level-crossing principles has nothing to do with hesitation, either in belief or action. If the principles were true, there should be cases where higher-order evidence, evidence about the nature of one's evidence or capacity, licences decisive belief or action that is not licensed by the first-order evidence. And once we see what such a licensing looks like in practice, the level-crossing principles look less attractive. So my main aim here is to expand the diet of examples that we have, and judge explanations by how well they handle all the examples in this class.

I already introduced one of the proposed examples in the previous chapter: David Christensen's example of the medical resident. I'm going to argue that the details of the case are underspecified in important ways. Once we fill in those details, it becomes clear that there are ways to respond to the case without thinking that they provide any support for level-crossing principles. Since we'll discuss the example at some length, it's worth repeating it here.

> I'm a medical resident who diagnoses patients and prescribed appropriate treatment. After diagnosing a particular patient's condition and prescribing certain medications, I'm informed by a nurse that I've been awake for 36 hours. Knowing what I do about people's propensities to make cognitive errors when sleep-deprived (or perhaps even knowing my own poor diagnostic track-record under such circumstances), I reduce my confidence in my diagnosis and prescription, pending a careful recheck of my thinking. (Christensen 2010a, 186).

First, a relatively trivial point. Many of the examples in the literature to date are written as either first-personal narratives, as this one is, or as second-personal narratives. It's not particularly easy to write commentary on such narratives. How, exactly, should we refer to the protagonist of the story? Should we call him David? That seems informal, and incorrect. I've been using the clumsy 'the narrator' or 'the resident', but those aren't the easiest phrases to track, especially over time. So it's better to give the protagonist a name. For similar reasons, it is better to say what exactly the diagnosis is, so we can easily refer back to it directly. There are two scope ambiguities in *David doubts that his diagnosis is supported by his evidence*, and those ambiguities can be cleared up if we specify what the diagnosis is, and what the evidence for it is.

While there are these general reasons to eschew first-personal narratives, there is an extra reason for concern here. The externalist thinks that is very important to distinguish evaluation of states from evaluations of agents, and to distinguish both of these from advice. We're interested here in what it would be rational for the resident to believe. That's distinct, at least in principle, from what a wise resident would believe in the circumstances. And both of those are distinct, again at least in principle, from what would be advisable for the resident to believe; i.e., from what advice we should give the resident about how to deal with such situations. Using first-personal, or second-personal, narratives in philosophical examples encourages conflation of rationality, wisdom and advisability. And we're wading into territory where it is important to remember those can come apart.

Returning to this example, Christensen does not make clear whether the doubts that have been raised are focussed in the first instance on the rationality of the resident, or on the reliability of the resident. (Indeed, the parenthetical remark seems to point in the opposite direction to the main text on just this point.) This

distinction may be important.[1] That is, it may be that the rational response to learning that one is prone to irrationality is very different to the rational response to learning that one is prone to unreliability. Maybe that won't be so, but at the beginning of inquiry there is little reason to think these two responses are certain to go together. So let's keep them apart in the examples we introduce.

I'm going to spend a lot of time on these three cases. All of them have a similar structure to Christensen's case, but with many more details filled in.

> Raisa is a medical resident with a new patient. He came in complaining of a burning sensation in his scalp and a nasty smell that he can't explain. Raisa looks at him and sees his hair is on fire. She decides that this is the cause of his symptoms, and starts to put the fire out. She is then told that she has been on duty for 36 hours, and that residents who have been on duty that long are typically over-confident in their diagnoses and prescriptions. What should she believe and do?

> Regina is a medical resident with a new patient. The whites of his eyes are yellow, and he is lethargic. Regina was taught in medical school that literally every lethargic patient with yellow eyes is jaundiced. (This is, we'll assume, actually true in Regina's world, though I'm sure it is actually false.) And she was taught, correctly, that every jaundiced patient should be treated with quinine. In her world, quinine cures all cases of jaundice and is, unlike every other medicine, free of all adverse side-effects. (Remember this is a fictional example!) So Regina prescribes quinine, recalling these facts from her medical training. But she is then told that she has been on duty 36 hours, and that residents who have been on duty that long are typically over-confident in their diagnoses. What should she believe and do?

> Riika is a medical resident with a new patient. He has a fever, headache, muscle and joint pains, and a rash that blanches when pressed. And he has recently returned from a trip to Louisiana. It seems to Riika that her patient has dengue fever, and that he should be treated with paracetemol and intravenous hydration. This is

[1] Indeed, in later work Christensen (2016) himself is very clear on the importance of this distinction, and what I say here draws on that later work.

> right; Riika's patient does actually have dengue fever, and it's rational to make that diagnosis after correctly processing the available evidence. But then Riika is told that she has been on duty for 36 hours, and that residents who have been on duty that long are typically over-confident in their diagnoses. What should she believe and do?

My judgment on these cases is that Raisa should keep trying to put out the fire, Riika should get a second opinion, and hold off on the treatment if it seems at all safe to do so, and that Regina's case is rather hard. That is, the details of what the symptoms are, and what the diagnosis and prescription are, matter to the judgment about what they should believe and do.

Now note that this doesn't immediately get Change Evidentialism off the hook. All it takes to refute Change Evidentialism is one case, and Riika's case may be enough to get the job done. But Raisa's case, and Regina's too, are important. Our best theory should explain what's true about those cases, and explain why the cases are different from Riika's. (If, indeed, Regina's case is different.) Ideally, they would even explain why Regina's case is a hard case, though maybe that's too much to ask of a philosophical theory (Ichikawa 2009).

As you may have guessed, I'm going to argue that Change Evidentialism does the best job at discharging these explanatory burdens. Before I start showing that, we need one more case. Christensen's example is one where the higher-order evidence seems to push in the direction of being more uncertain. All of the cases from the literature that I cited earlier have the same feature. But in principle we can imagine cases that go the other way.

> Roshni is a medical resident with a new patient. His symptoms are similar to those of Riika's patient, but his rash does not blanch when pressed, and indeed is light enough that it doesn't have the distinctive visual characteristics of the rash produced by dengue fever. Given his symptoms and history, Roshni thinks he probably has dengue fever, though the oddity of the symptoms means that she thinks other diagnoses are possible. So she wants to run more tests before commiting to any course of treatment. One reason for her to run more tests is that she remembers there are some other illnesses going around that display similar symptoms to what her patient displays. Roshni is then told that she has been on duty for 13 hours and (and this is actually true in the world of the story) that residents who have been on duty between 12 and 14 hours are

> typically over cautious in their diagnoses. If such a resident thinks probably p, then p is almost always true, and the resident should simply have come to believe p. Now as it turns out Roshni is an exception to this rule; she really doesn't have strong enough evidence to conclude that her patient has dengue fever, and she's right to stop at the conclusion that he probably has dengue. But she has no independent reason to believe that she is an exception. So what should she believe and do?

It would be wrong for Roshni to reason as follows.

> When someone in my circumstance concludes probably p, then there is almost always sufficient evidence to conclude definitely p. I've concluded he probably has dengue fever. So he definitely has dengue fever. So I'll stop running tests and start the prescribed treatment for dengue fever.

Roshni can't rule out other possible diagnoses simply on the basis of general characteristics of residents in situations like her. If her patient has some other disease, and Roshni treated him for dengue on the basis of higher-order considerations, she'd be guilty of malpractice.

So now we have another task for our theory to perform. It must explain why there is, to use a term Stewart Cohen suggested to me, *epistemic gravity*. Riika's case shows that, at least sometimes, intuition wants agents to lower confidence when they learn they are in a situation where people are often over-confident. But Roshni's case shows that the converse is not always true. Higher order evidence can, according to intuition, make confidence go down but not up. And that's especially true if one had judged correctly to begin with.

I'm going to argue that a theory that rejects level-crossing principles, and accepts Change Evidentialism, is best placed to explain these four cases.

8.2. Diagnoses and Alternatives

It is easy to see why one might think Riika's case is a problem for Change Evidentialism. Imagine that Riika's twin sister is also a medical resident, and looks at the same public data about Riika's patient. And she, like Riika, concludes that the patient has dengue fever. Now the residents are both told that Riika (but not

her sister) has been awake for 36 hours, and hence a member of a class that is systematically over-confident in their diagnoses. This seems like a reason for Riika, but not her sister, to reduce their confidence that the patient has dengue fever. And that's a problem for Change Evidentialism. That Riika has been awake for 36 hours either is, or is not, evidence against the hypothesis that the patient has dengue fever. If it is, then both sisters should become less confident. If, more plausibly, it is not, then if Riika should change, that violates Change Evidentialism.

There is a purely technical solution to this problem that I mention largely to set aside. The argument of the previous paragraph assumed that when the nurse told Riika how long she'd been awake, the evidence Riika received was a proposition like *Riika has been awake for 36 hours*. That's evidence that Riika can get, and that her sister can get. And intuitively learning that has a different effect on the two of them. But we could conceptualise Riika's evidence differently. We could think her evidence is a centered world proposition, in the sense popularised by David Lewis (1979). On this picture, Riika's evidence is *I have been awake for 36 hours*, while her sister's evidence is *My sister has been awake for 36 hours*. So they get different evidence. So there is no argument that Change Evidentialism fails.

This feels a bit like a cheat, at best. After all, we can imagine that the nurse explicitly says to the pair of them, "Riika has been awake for 36 hours". In that case it would feel extremely artificial to say that the evidence is really this first-personal claim about Riika. But while this technical attempt to save the letter of Change Evidentialism isn't attractive, it tells us something useful. The information about Riika's sleep (or lack thereof) matters to Riika because of what it tells her about her mind, i.e., about the very mind she is both using to think about the patient, and thinking about. And an explanation of what goes on in the case should be sensitive to this fact.

It is important that Riika and her sister are medical residents. The patient in the next bed can't reasonably believe that Riika's patient has dengue fever on the basis of the data. Or at least he can't unless he has medical training. Should we think this is a case where different people with the same evidence can draw different conclusions? No, because this data about the patient does not exhaust the evidence. The evidence also includes everything relevant that Riika learned in her medical training. That's evidence she has in common with her sister, but not with the patient in the next bed.

The evidence provided by training, and background information, has to play two

roles. First, it has to make it plausible that the patient has dengue fever. It does that by including facts about the symptoms the patient displays, and facts about what symptoms patients with dengue fever typically display. But it must also play a second role. In making a diagnosis and a prescription, Riika isn't just saying that the patient has dengue fever. She is also saying that dengue fever is the cause of the symptoms. And that requires excluding a lot of other possible diseases, either on the basis that they are inconsistent with the symptoms displayed, or because they are initially implausible and the evidence does not sufficiently raise their likelihood to make them worth taking seriously. If the patient has dengue fever and some other equally serious disease that causes some of the symptoms, then to diagnose dengue fever is to some extent to mis-diagnose the patient. And to start the treatments for dengue fever is, in such a case, to mis-treat the patient. In these respects, forming a diagnosis of dengue fever is importantly different, and stronger, to forming a belief that the patient has dengue fever.

This exclusion of alternative diseases must be prior to the diagnosis of dengue fever. Imagine how strange it would sound for Riika to have this conversation with her supervisor:

> Supervisor: Why do you think that the patient does not have West Nile?
> Riika: Well, the patient has a fever, headache, a rash etc.
> Supervisor: Yes, those are all consistent with West Nile.
> Riika: Ah, but you see, from those symptoms we can conclude that the patient has dengue fever.
> Supervisor: Yes, and?
> Riika: So the symptoms have been fully explained, so there is no reason to believe the patient has West Nile.

That's not good reasoning. It would be perfectly good to reason that the symptoms aren't consistent with West Nile, so the patient doesn't have West Nile. Or that West Nile is very rare among people with the patient's background, so it is better to conclude that he has a disease that is (much) more prevalent in areas he has been. But it isn't good to first diagnose the patient with dengue fever, and use that to conclude they don't have West Nile.

So a good diagnosis draws on lots of background information. So that information must in some sense be available to the doctor. I don't mean that the information has to be accessible in the sense that she could recite it off hand. But she must be able to base her diagnosis on the background information. And if she's been awake for 36 hours, then that information is probably not available,

even in this weak sense. As I will discuss in section 8.4, there are hard questions about just when it is that evidence previously acquired can still be used. But it is plausible that the relevant information that excludes other diagnoses is not something Riika can use in her tired state.

There is another complication to consider here. Riika has to rule out particular alternatives like West Nile before she can diagnose the patient with dengue fever. But she also has to rule out, collectively, alternative explanations she hasn't thought of, or may have forgotten. It's not enough that the alternative explanations simply fail to exist. If one knows the patient has yellow eyes, and as a matter of fact the only possible explanation for this is that they are jaundiced, it doesn't follow that one is in a position to rationally conclude the patient is jaundiced. One must know that only jaundice causes yellow eyes, or at least that it's the only plausible cause. And the same holds for all other diagnoses.

It is here that concerns about one's own alertness become particularly pressing. At least in my own case, the most worrying consequence of excessive tiredness is that I overlook alternative explanations of phenomena. When that happens, my abductive inferences to particular explanations are unreasonable because I should have looked harder for alternatives before settling on one explanation. So let's spend some time thinking about how this might affect the reasonableness of Riika's diagnosis.

8.3. Tiredness and Abduction

We'd like to show that NR is true, and even better, that LNR is true, without positing any kind of level-crossing principle.

NR It is Not Reasonable for Riika to believe that her patient has dengue fever.

LNR When she Learned that she had been awake for 36 hours, it became Not Reasonable for Riika to believe that her patient has dengue fever.

Since we're not using level crossing principles, we can't reason as follows.

1. Riika has been awake for 36 hours, and she knows this.
2. So it is reasonable for her to believe that her diagnoses are unreasonable.
3. Whenever it is reasonable to believe that some mental state is unreasonable, it is unreasonable to maintain that mental state.
4. It was not unreasonable to believe that she'd made a reasonable diagnosis before learning how long she'd been awake.

5. So, from 3 and 4, LNR is true.

If we want to reject level-crossing principles, then we have to reject step 3 of that purported explanation. We need to find something to put in its place. I'm going to offer three explanations. The first two are probably flawed. But I'm offering them in part because they aren't obviously wrong, and would solve the problem without appeal to level-crossing principles. And, more importantly, thinking through what's wrong with these explanations helps us see what's right about the correct explanation of Riika's case. Here is the first of these probably flawed explanations.

1. To reasonably conclude that *p* by abductive inference, Riika needs to antecedently, reasonably believe that other explanations of the data fail.
2. Her best evidence is that other explanations of the data fail is that (a) it seems to her that no other explanation works, and (b) she is a reliable judge of when alternative explanations are available.
3. When she learns she has been awake for 36 hours, she is no longer in a position to reasonably use part (b) of that evidence.
4. So LNR is true; once she learns that she has been awake for 36 hours, she can no longer reasonably make the abductive inference from the data to the diagnosis of dengue fever.

I suspect there are two, related, mistakes in this explanation. It relies on a 'psychologised' conception of evidence, and Timothy Williamson (2007) has argued convincingly against that conception of evidence. It isn't at all obvious that Riika has to reason from how things seem to her to conclusions about the world in order to form medical diagnoses.

And it isn't obvious that Riika's has to form a reasonable belief that there are no alternative explanations, and that she has to do so before forming the diagnosis. It might be that an abductive inference is reasonable if one's evidence rules out alternative explanations of the data, and one is reliably disposed to consider alternative explanations when they are not ruled out. In other words, an abductive inference might be good (in part) in virtue of being based in a skill in considering explanations, and that skill may be manifest when the abductive conclusion is drawn, not antecedently to it being drawn.

Even if all that is true, there is still a skill that is needed. That skill needs to reliably rule out alternative explanations. And Riika is really tired; maybe she can't exercise that skill while so tired. This idea leads to our second (probably mistaken) explanation.

1. To reasonably conclude that *p* by abductive inference, Riika has to be able to reliably rule out alternative explanations as unreasonable.
2. Since she's been awake for 36 hours, Riika cannot reliably rule out alternative explanations of the symptoms as unreasonable.
3. So NR is true; Riika cannot make the diagosis reasonably because she cannot reliably rule out alternatives.

One shortcoming of this explanation is that it doesn't explain LNR. Indeed, if the premises here are true, then LNR is in fact false. It is the fact that Riika has been awake for 36 hours that makes her diagnosis unreasonable, not her learning that she's been awake that long. To the extent that we think LNR is true, that's a reason to dislike the explanation.

A bigger problem for this explanation is that we don't really know that premise 2 is true. What we know is that folks in general who have been awake as long as Riika as not reliable. But perhaps she is an exception. Indeed, the setup of the example suggests she may well be an exception. The fact that other people in her position are unreliable does not entail that she is unreliable. Or, at least, it doesn't entail this without some strong assumptions about the reference class that is relevant to Riika's reliability. So let's try a different explanation.

The alternative explanation starts with the observation that the reliability of a mechanism is not normally enough for it to produce reasonable, or rational, beliefs. If a scale is working, but there is excellent testimonial evidence that it is not working, it is unreasonable to believe what the scale says. This applies to internal mechanisms too. If one is reliably told that one is in an environment full of visual illusions, it is unreasonable to believe what one sees, even if one's eyesight is reliable.

A similar story holds true for skills. To learn that the patient has dengue fever, Riika has to exercise her skill at reliably ruling out alternative explanations of the data. And while she has such a skill, she has no reason to believe that she has it. Indeed, she has a positive reason to believe that she lacks it, since she has been awake so long, and people who have been awake that long typically lack the skill. So she should not rely on the skill. Here, then, is my preferred explanation for what's going on in Riika's case. I'll call that explanation the *evidentialist explanation* in what follows, since it makes key use of how evidence does (or does not) change in explaining changes in what states it is rational to hold.

1. To reasonably conclude *p* by abductive inference, Riika must reasonably rely on her skill at excluding alternative explanations of the data.

2. It is not reasonable to rely on a skill if one has excellent, undefeated, evidence that one does not currently possess the skill.
3. So, once Riika learns she has been awake 36 hours, she cannot reasonably infer from the observed data to the conclusion that the patient has dengue fever.

If this explanation is correct, the case is not a counterexample to Change Evidentialism, and we do not need to appeal to level-crossing principles. Riika had to rely on her sensitivity to explanations she had not considered in order to have a justified diagnosis. Even though she is, in the circumstances, sufficiently sensitive to alternative explanations, she could not reasonably rely on that sensitivity when she has such good evidence that her skills are temporarily diminished. So her belief that the patient has dengue fever is unjustified.

That is our explanation of why Riika loses knowledge, and loses reasonable belief, when she learns that she has been awake for 36 hours. But it isn't the only possible explanation. There are, for example, explanations that appeal to level-crossing principles. Why should we prefer the explanation I just offered? As I'll argue in the next section, the answer is that only this explanation in terms of skill can generalise to cover all of the cases.

8.4. Explaining all Four Cases

Let's start with Raina. Unlike Riika, Raina needs neither specialist background information, nor expert insight, to form a diagnosis. There's a guy with his hair on fire, and she comes to the belief that his hair is on fire. She perhaps needs the background information that burning hair burns and smells, and has a distinctive fiery appearance, but most adults will have that information ready to hand in case of emergency. So the kinds of evidence that are threatened by fatigue are not needed to form the judgment in Raina's case. So she still knows, even in her fatigued state, that her patient's hair is on fire. Since judging that the patient's hair is on fire doesn't require any particular skill, it doesn't matter that her skills are diminished.

Unlike Riika, Roshni didn't have enough public information to conclude her patient had dengue fever. She needed the extra step that there are no other plausible explanations of the data. Since there are other plausible explanations of the data, she can't know there are none. Hence it cannot be part of her evidence that there are none. Being fatigued might explain why one's 'insights' do not really

constitute evidence. But it can't turn non-insights, and non-facts, into evidence. So even in her semi-fatigued state, Roshni still lacks sufficient evidence to diagnose her patient with dengue fever. So she still doesn't know her patient has dengue fever, as we hoped to explain.

We'll spend much more time on this in chapter 11, but for now note one quick reason to suspect that Roshni's credence that her patient has dengue fever should not move at all. Assume that she learns not just that residents who have been on duty 12–14 hours are systematically under-confident in their diagnoses, but that they remain so after making their best efforts to incorporate this information about their own under-confidence. And assume that Roshni should, on learning that she is part of a group that is systematically under-confidence, increase her confidence in her preferred diagnosis. Now we have a perpetual confidence increasing machine. Even once she has increased her confidence in light of the information about herself, she has reason to increase it again, since she is still in a group that systematically is too cautious in their judgments. And this fact persists no matter how hard she tries. But perpetual confidence increasing machines, like perpetual motion machines, are absurd. The best place to stop this machine is at the very start. So Roshni should not increase her confidence at all. (I think this is intuitively the right thing to say about her case, but this argument is offered to those who don't share the intuition.) And that in turn provides reason to not just believe the evidentialist explanation of Riika's case, but to believe the 'non-psychologised' version of that explanation.

The really tricky case, from this perspective, is Regina. She doesn't need any skill in identifying possible alternative explanations of the data. She just needs to remember some facts from her medical training, make some straightforward observations, and perform a very simple logical deduction. Her tiredness does not affect her ability to make the observations or, I suspect, to do this deduction. A tired person may struggle to draw complicated consequences from data, but going from *All Fs are Gs* and *This is F* to *This is G* does not require particular skill.

The big question is whether Regina can really rely on her memory when she is tired. It is helpful to think about this case by comparing it to the Shangri-La example developed by Frank Arntzenius (2003). Here is the slightly simplified version of the case that Michael Titelbaum sets out.

> You have reached a fork in the road to Shangri La. The guardians of the tower will flip a fair coin to determine your path. If it comes up heads, you will travel the Path by the Mountains; if it comes up tails,

> you will travel the Path by the Sea. Once you reach Shangri La, if you have traveled the Path by the Sea the guardians will alter your memory so you remember having traveled the Path by the Mountains. If you travel the Path by the Mountains they will leave your memory intact. Either way, once in Shangri La you will remember having traveled the Path by the Mountains. The guardians explain this entire arrangement to you, you believe their words with certainty, they flip the coin, and you follow your path. What does ideal rationality require of your degree of belief in heads once you reach Shangri La. (Titelbaum 2014, 120)

The name of the person Titelbaum's narrator is addressing isn't given, so we'll call him Hugh. And we'll focus on the case where Hugh actually travels by the Mountains.

There is something very puzzling about Hugh's case. On the one hand many philosophers (including Arntzenius and Titelbaum) report a strong intuition that once in Shangri La, Hugh should have equal confidence that he came by the mountains as that he came by the sea. On the other hand, it's hard to tell a dynamic story that makes sense of that. When he is on the Path by the Mountans, Hugh clearly knows that he is on that path. It isn't part of the story that the paths are so confusingly marked that it is hard to tell which one one is on. Then Hugh gets to Shangri La and, well, nothing happens. The most straightforward dynamic story about Hugh's credences would suggest that, unless something happens, he should simply retain his certainty that he was on the Path by the Mountains.

Resolving the tension here requires offering a theory of the epistemology of memory. And I have no desire to do that, any more than I had a desire in the ethics part of the book to offer a first-order ethical theory. What I am going to do is say why hard questions within the epistemology of memory are relevant to what we should say about Hugh's case, and by extension Regina's case.

Some theories of memory are synchronic. Whether the agent's mental state at time t makes it rational for her to believe that p, on the basis of her (apparent) memories, solely depends on the the properties she possesses at t. There are two natural ways to fill in the synchronic theory. First, we could say that the agent's faculty of memory outputs propositions that become, if it is a reliable faculty, evidence for the agent. (It's presumably a gross oversimplification of the best cognitive and neural theories of how memory works in humans to describe it as a faculty, but we'll have to work with such simplifications to get a broad enough

view of the philosophical landscape.) Second, we could say that the apparent memories the agent has provide her evidence, and she can then reason using either what she knows about herself, or perhaps some default entitlements to trust herself that she possesses, to the truth of the contents of those memories.

On either kind of synchronic theory, Hugh won't know that he came to Shangri La via the mountains. If memory provides evidence directly, it does so only when it is reliable. And on this question, it is unreliable, since in nearby worlds it produces mistaken outputs. It's true that there is nothing funky about the causal chain leading to Hugh's memory. But on a synchronic theory of memory, the nature of the chain is not relevant; all that is relevant is the reliability of the output. And the output is not reliable. If, on the other hand, the evidence is something like the apparent memory Hugh has, then things are even worse. He knows that he can't reason from his apparent memory to any claim about how he got to Shangri La, because in very nearby worlds his apparent memories are badly mistaken.

Arntzenius argues that Hugh should have a credence of 0.5 that he came by the mountains as follows. (Assume Arntzenius is talking to Hugh here, so 'you' picks out Hugh.)

> For you will know that he would have had the memories that you have either way, and hence you know that the only relevant information that you have is that the coin was fair. (Arntzenius 2003, 356)

That argument seems to presuppose that we are using the second, psychologised, version of the synchronic theory of memory. If we understand memories to be not just phenomenal appearances, but traces of lived experiences, then Hugh would very much not have the memories that he has either way. He might think that he had the same memories had he come by the sea, but he'd be wrong. Still, Arntzenius's argument doesn't seem to rely on this feature of memory. What it does seem to rely on is that in an important sense, Hugh would be the same right now however he had arrived at Shangri La. That is, it relies on a synchronic theory of memory. Sarah Moss (2012) makes a similar claim about the case. (Again, her narration is addressed to Hugh.)

> Intuitively, even if you travel on the mountain path, you should have .5 credence when you gets to Shangri La that the coin landed heads. This is a case of abnormal updating: once you arrive in Shangri La, you can no longer be sure that you traveled on the

> mountain path, because you can no longer trust your apparent memory. (Moss 2012, 241–42)

Again, the presupposition is not just that we have a synchronic epistemology of memory, but that the evidence memory provides comes from appearances. And, once again, the second presupposition does not seem to really matter. We would get the same result if we took memory to provide evidence directly, but only when it was reliable. What matters, that is, is the synchronic epistemology of memory.

In recent work, Moss (2015) has developed a systematic defence of synchronic epistemology, what she usefully calls 'time-slice epistemology'. And while she makes a good case for it, there is also a good case for a diachronic epistemology. Richard (Holton 1999, 2014) has argued for diachronic norms of intention, and for understanding belief as being in important ways like intention. From these premises he concludes that there are diachronic norms on belief. David James Barnett (2015) has offered more direct arguments for adopting a diachronic epistemology of memory. So we should work through what happens in cases like Shangri La on a diachronic approach.

It turns out that we quickly face another choice point. The cases we are interested in are ones where an agent knows *p* at an earlier time t_1, and then this belief is preserved from t_1 to a later time t_2. The theoretical choice to make is, is this sufficient for the agent to know *p* at t_2, or could the knowledge be defeated by things that happen in the interim? If the knowledge could not be defeated, then Hugh knows he came by the mountains, for the obvious reason that he once knew this and has never forgotten it. If it can be defeated, then on any of the most obvious ways to incorporate defeat into the theory, Hugh's claim to knowledge will be defeated. He is, after all, part of a group (explorers who arrive at Shangri La) who have very unreliable memories, and he knows that.

Whatever we say about defeat here can be made consistent with Change Evidentialism[2]. Since we're developing a diachronic epistemology, we should allow that evidence can be accrued over time. On the version of the theory where memories are indefeasible, Hugh's evidence that he came via the mountains is his perception of the mountain path. This perception can be his evidence well into the future, as long as his memory does its job of preserving the visual evidence. (He could of course forget how he got to Shangri La, but we're only discussing cases where beliefs are preserved throughout the relevant time period.)

[2]Note that the key notion in the statement of Change Evidentialism is *change* of evidence, not accrual of evidence. Losing evidence matters too.

If memories can be defeated, the Change Evidentialist should say that the defeaters prevent the past perceptions from being current evidence. (In general, I think the evidentialist should say that defeaters prevent propositions becoming part of one's evidence. But defending that claim would take us too far afield.) If his evidence does include the contents of his perceptions while on the path, then he now knows that he came via the mountains, if it does not he does not. Either way, it is the change or lack of change of evidence (and not merely his worries about his own reliability) that explain why he knows what he does.

I've described four theories of memory, two synchronic and two diachronic. On three of the four theories, Hugh does not know, indeed does not even have reason to be particularly confident, that he came by the mountain. On the fourth he does know this. I think that's a reasonable stopping point; it's left as a somewhat difficult philosophical question whether Hugh knows that he came via the mountains. But it's not one we have to settle the big picture views I've been defending, since either answer to the philosophical question about memory is consistent with those views.

And what we say about Hugh carries over to Regina's case. The big issue is whether she (still) has the following two propositions as evidence.

1. All lethargic patients with yellow eyes are jaundiced.
2. All jaundiced patients should be treated with quinine.

If she has 1 and 2, then she should treat her patient with quinine. This isn't, or at least isn't just, because 1 and 2 entail that she should treat her patient with quinine. It's rather because these pieces of evidence provide strong and immediate support for the claim that she should treat her patient with quinine.

Does she (still) have those propositions as evidence, or as something she can derive and use as evidence? On either synchronic theory of memory, she does not. Her apparent memory of 1 and 2 cannot ground an inference to the truth of 1 and 2, since she knows that she is unreliable given her fatigue. Alternatively, if memory delivers propositions like 1 and 2 directly, the fact that she is so fatigued right now will defeat memory's claim to being a source of evidence. If we adopt a diachronic theory of memory, then what matters is whether we allow for (anything like) defeaters. If we do, her current fatigue is, probably, a defeater, so she again doesn't know that her patient should be treated with quinine. But on the (not totally implausible!) diachronic theory that rejects defeaters, we get that

she does know. I think this is the right result; Regina's case is not as clear as Riika's, and it is right that it turns on hard philosophical questions.[3]

If we explain Riika's case using level-crossing principles, then we should say that Regina's case does not turn on hard philosophical questions. On this approach, Regina's case is easy. She can't rationally believe that she rationally believes that the patient is jaundiced, so she can't rationally believe that the patient is jaundiced. Now this seems to me to be the wrong result in Regina's case. It's wrong twice over; it says the wrong thing about Regina, and it says the case is easy when in fact it is hard. But because the question is hard, I don't want to lean any argumentative weight on it. And I doubt that we should ever put much argumentative weight on intuitions about whether cases are hard or easy. Instead I'll argue against the application of level-crossing principles to Riika's case by comparing Riika's case with Raina's and Roshni's.

The level-crossing explanation of Riika's case provides no resources to distinguish between Riika's case and Raina's. Both of them have reasonably responded to the evidence that is available. Both of them then get evidence that they are (temporarily) unlikely to be responding correctly to evidence. These facts are, in Riika's case, held to be sufficient to explain why she should change her view. But they are features of Riika's case that are shared with Raina's case. Since Raina should not change her view on being told she has been awake for 36 hours, we need either something more, or something else. An explanation of Riika's case based on level-crossing principles will over-generalise; it will 'explain' why Raina should change her mind too.

Roshni is even more of a challenge for explanations that rely on level-crossing principles. Let p be the proposition *The patient might not have dengue fever*. At the start of the story, Roshni believes that, and rationally so. But then she gets evidence that she cannot rationally form beliefs like that given her state. So, if the level-crossing principle is true, then she should lose the belief in p. But if she thinks that the patient's having dengue fever is at least very likely, and does not believe that it might be false, that sounds to me like she believes it. That is, the only way to comply with the level-crossing principles is to believe the patient does have dengue fever. And that conclusion is absurd.

So Roshni is a counterexample to a lot of level-crossing principles. The following claims about her are true:

[3] In a recent paper (Weatherson 2015) I take a stand on some of these questions about memory in ways that go beyond what is necessary for rejecting level-crossing.

- Roshni rationally believes that *p*.
- Roshni could not rationally believe that she rationally believes that *p*.
- Roshni should believe that her evidence does not support rational belief in *p*.

And level-crossing principles are meant to rule out just those combinations. So Roshni's case does not just undermine an abductive argument for level-crossing principles, it provides direct evidence that those principles are mistaken.

8.5. Against Bracketing

David Christensen draws a different response to these puzzles involving higher-order evidence. His theory is that higher-order evidence requires us to 'bracket' first-order evidence. Here is how he introduces the idea. (The background is that he is discussing a case where he did a logic problem, got the right answer, and then was told he took a drug that distorts most people's logical abilities.)

> It seems to me that the answer comes to something like this: In accounting for the HOE (higher order evidence) about the drug, I must in some sense, and to at least some extent, *put aside* or *bracket* my original reasons for my answer. In a sense, I am barred from giving a certain part of my evidence its due. After all, if I could give all my evidence its due, it would be rational for me to be extremely confident of my answer, even knowing that I'd been drugged. In fact, it seems that I would even have to be rational in having high confidence that I was immune to the drug. By assumption, the drug will very likely cause me to reach the wrong answer to the puzzle if I'm susceptible to it, and I'm highly confident that my answer is correct. Yet it seems intuitively that it would be highly irrational for me to be confident in this case that I was one of the lucky immune ones. ... Thus it seems to me that although I have conclusive evidence for the correctness of my answer, I must (at least to some extent) bracket the reasons this evidence provides, if I am to react reasonably to the evidence that I've been drugged. (Christensen 2010a, 194–95, emphasis in original)

There are a few different arguments here that we need to tease apart.

There is an argument that bracketing is needed because otherwise the narrator will have 'conclusive' evidence for the answer to the logic problem. This isn't right; or at least it is misleading. In a sense seeing my coffee cup on my desk is conclusive evidence for the truth of any mathematical proposition. It does entail it. But it's a terrible reason to believe, for example, Fermat's Last Theorem. There is another sense of conclusive that is more relevant; whether some evidence provides epistemically conclusive reason to believe a conclusion. And mere entailment does not suffice for that.

There is an argument I think implicit in Christensen's remarks that if we allowed the first order evidence to stand, we'd be licencing some improperly circular reasoning. That's an interesting observation, and I'll discuss it at more length in the next chapter.

But what we're interested in is the conclusion, that the original evidence must be bracketed or set aside in cases where higher order evidence suggests we are likely to be making a mistake. And that conclusion, we can now see, can't be right. It can't be right because of Raina's case and Roshni's case. If Raina brackets her first order evidence, she won't have reason to put out the fire in her patient's hair. But she has excellent, indeed compelling, reason to do that. And if Roshni brackets her first order evidence, she will have sufficient reason to believe that her patient has dengue fever, and to start treating him. But she does not have sufficient reason to do that.

These cases aren't isolated incidents. They point to two general problems with the bracketing picture. It doesn't distinguish between cases where evidence immediately supports a conclusion, and cases where the evidence supports the conclusion more indirectly. The latter cases, ones where the agent must use the initial evidence to derive more evidence, and then use the larger evidence set to support the conclusion, are cases where higher order evidence matters. But the reason higher-order evidence matters in those cases is that higher-order evidence blocks those intermediate steps. Cases like Raina's are different, but the bracketing story does not distinguish them. And the bracketing story can't explain the existence of epistemic gravity, while the evidentialist explanation I've offered can.

There are other cases that, while not as clear, seem to me cases the bracketing story cannot handle correctly. The following case is inspired by some examples presented Jonathan Weisberg (2010).

Jaga has been taking some medication. She knows that she has taken the medication for 22 days, and that she has taken 18 pills each day. She then learns some

very worrying news. The medication is being withdrawn from sale because it has a striking effect on anyone who takes 400 or more pills; it makes them incredibly bad at arithmetic for several weeks. The effect is surprisingly sharp in its effect; anyone who has taken 399 or fewer is unaffected, but once one has taken the 400th pill, it kicks in with full force. (Yes, this is a very unrealistic case, but more realistic cases are possible, and would simply be more complicated to discuss.)

Now Jaga is very worried. She knows that she has taken 22 times 18 pills. But she is unsure what 22 times 18 is. That's not unreasonable; most of us wouldn't know what it is off the top of our heads either, without doing the calculation. And one of the things that worries Jaga is that before doing the calculation, it seems pretty likely to her that it is greater than 400. And that isn't unreasonable either. It's wrong, but well within the reasonable range of error.

So Jaga does the calculation. She works out that 22 times 18 is 20 times 18 plus 2 times 18, so it is 360 plus 36, so it is 396. Wonderful, she thinks, I haven't taken too many pills. So I can do arithmetic well, as indeed I just did. That's exactly the right attitude for Jaga to have. Her evidence does not actually show that she is bad at arithmetic. Before she sat down to do the calculations, she should have worried that she was bad at arithmetic. But now that she's done the calculations, she knows better.

But note this isn't what a defender of the bracketing view can say about Jaga's case. There is a serious doubt about whether she is good at arithmetic, and relatedly about whether she has taken 400 or more pills. She can't resolve that by appeal to her first order evidence about whether she has taken 400 or more pills, since whether her calculations provide her with reason to believe that she's taken 400 or more pills is exactly what is at issue. More formally, let p be the proposition that she's taken less than 400, and q be the proposition that she's good at arithmetic. The intuition behind the bracketing view is that one can't come to believe q by doing some arithmetic and trusting your answers. Yet that is exactly what Jaga has done, admittedly via the roundabout route of coming to believe p, and antecedently knowing that q is true iff p is true.

The point of Jaga's case is that bracketing has implications not just in cases where an agent gets evidence that does suggest she is irrational or unreliable, but also in cases where she gets evidence that might suggest that. And those implications are much less plausible than they are in the cases where the force and direction of the higher-order evidence is clearer. We'll return to such cases extensively in chapter 10. The next priority, however, is to deal with the circularity worry.

If we reject level-crossing principles, and accept Change Evidentialism, are we committed to accepting what are in fact bad kinds of circular reasoning?

9. Circles, Epistemic and Benign

9.1. Normative Externalism and Circularity

Some of the views that I'm opposing are motivated by anti-circularity considerations. Consider, for instance, the principle David Christensen calls Independence, which is a version of the bracketing principle that was the focus of the previous section. I'm quoting it here with the argument for it that immediately follows.

> **Independence** In evaluating the epistemic credentials of another's expressed belief about P, in order to determine how (or whether) to modify my own belief about P, I should do so in a way that doesn't rely on the reasoning behind my initial belief about P.
>
> The motivation behind the principle is obvious: it's intended to prevent blatantly question-begging dismissals of the evidence provided by the disagreement of others. It attempts to capture what would be wrong with a P-believer saying, e.g., "Well, so-and-so disagrees with me about P. But since P is true, she's wrong about P. So however reliable she may generally be, I needn't take her disagreement about P as any reason at all to question my belief." (Christensen 2011, 1–2)

To my eyes, this argument seems to involve a category mistake. Moves in a dialectic can be question-begging or not. But here Christensen seems to want to put restrictions on rational judgments on the grounds that the alternative would be question-begging. That seems like the wrong way to get the desired end. If we want to stop "blatantly question-begging dismissals" we can just remind people not to be rude.

I think the problem Christensen is highlighting is not to do with question-begging, but to do with circularity. The problem is that if we violate Independence, we can use our reasoning to conclude that our reasoning is reliable,

and that's circular. Or, to be more accurate, it has a whiff of circularity about it. Trying to turn this into an argument for Independence though will be difficult.

Part of the difficulty is that it isn't easy to say exactly what the circularity involved is. Consider the following little example, where Chiyoko and Aspasia are discussing arithmetic. They know that exactly one of them has taken a drug that makes people bad at simple arithmetic. Chiyoko does some sums in her head, listens to Aspasia, and reasons as follows.

1. 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
2. Aspasia believes that 2+2=5, and 3+3=7, and 4+5=8, and 7+9=15, while I believe that 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
3. So, she got those four sums wrong, and I got them right.
4. It is likely that I would get at least one of them wrong if I'd taken the drug, and unlikely that she would get all four wrong unless she'd taken the drug.
5. So, probably, I have not taken the drug, and she has.
6. So I should not modify my beliefs about arithmetic in light of what Aspasia says; she has taken a drug that makes her unreliable.

It isn't clear to me just which step is meant to be circular. If Chiyoko had reasoned as follows, I could see how we might take her reasoning to be circular.

1. It seems to me that 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
2. It seems to Aspasia that 2+2=5, and 3+3=7, and 4+5=8, and 7+9=15.
3. From 1, it's true that 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
4. From 1, 2 and 3, my arithmetic seemings are reliable, and Aspasia's are not.
5. So, probably, I have not taken the drug, and she has.
6. So I should not modify my beliefs about arithmetic in light of what Aspasia says; she has taken a drug that makes her unreliable.

If Chiyoko reasons this way, the only reason for thinking she is right and Aspasia is wrong is her own judgment, which is exactly what is at issue in 6. But that isn't at all how people usually reason. Nor is it a sensible rational reconstruction of their reasoning. Rather, the first version of the inference is much more like the way normal human beings do, and should, reason. And in this case the symmetry of the dispute between Chiyoko and Aspasia is broken by a fact recorded at line 1, namely that 2 plus 2 really is 4, 3 plus 3 really is 6, and so on. And while Chiyoko uses her mathematical competence to come ot know that fact, she doesn't learn it by reasoning about her mathematical competence.

If she did, it would be a posteriori knowledge, whereas in fact it is a priori knowledge. So if there is some circular reasoning going on in the first inference, the circularity is fairly subtle, and it won't be easy to say just what it is.[1]

Still, there is some vague feeling of circularity that goes along with even that first inference. And in principle we shouldn't say that some reasoning is acceptable just because we can't precisely articulate the sin it commits. Compare: We shouldn't say that the Dharmottara cases described by Jennifer Nagel (2014, 57) are cases of knowledge just because it is hard to say exactly what makes them not knowledge.[2] Call this the 'whiff of circularity' objection to normative externalism, since normative externalism arguably licences the first form of reasoning, but there is a whiff of circularity about it. The aim of this chapter is to respond to the whiff of circularity objection. Much of our time will be spent trying to make the objection more precise. (As David Lewis almost said, I cannot reply to a whiff.) We'll start with the worry that the objection trades on a fundamental confusion between inference and implication.

9.2. Inference, Implication and Transmission

As Gilbert G. Harman (1986) has pointed out, it is very important to separate the theory of implication, i.e., logic, from the theory of inference, which sits in the intersection between psychology and epistemology. The following argument is perfectly valid, even though following it would make a lousy inference.

1. The Eiffel Tower is large.
2. The Eiffel Tower is not large.
3. So, London is pretty.

Using terminology drawn from work by Crispin Wright (2000, 2002), we might say this is a case where warrant does not transmit from the premises to the conclusion. An agent could not gain warrant for the conclusion of this argument by gaining warrant for its premises. But that does not tell against the validity of the argument. Whenever the premises are true, so is the conclusion. Any proof of

[1] David James Barnett (2014) also notes that it is important to distinguish the case where Chiyoko uses her mental faculties from the case where she reasons about them. He thinks, and I agree, that once we attend to this distinction, it is far from clear that there is anything problematically circular about what Chiyoko does.

[2] Cases with the same structure as Dharmottara's became the focus of some discussion in the Anglophone philosophical tradition after they were independently discovered by Edmund Gettier (1963).

the premises can be converted into a proof of the conclusion. And so we have excellent reason to believe the argument is valid, even though it does not ground any good inference.

Rather than using Wright's slightly technical term 'warrant', we'll focus on the class of Potential Teaching Arguments, or PTAs. These are arguments where an agent could come to learn the conclusion by first learning the premises, and then reasoning from them to the conclusion. The modal term 'could' there is context-sensitive, and vague. The context sensitivity comes from the fact that whether an argument is a PTA might depend on which agent we are focussing on, and on how that agent came to know the premises. Imagine, for example, that Marie is a scientist who is working on a machine to measure the relative radioactivity of two substances. The machine is, it turns out, very accurate, but it is also the first of its kind, and the theory behind it is somewhat speculative. Now consider this argument.

1. Marie's machine says that a is more radioactive than b.
2. In fact, a is more radioactive than b.
3. So, Marie's machine is accurate about a and b.

That's a valid argument, but it isn't a PTA. At least, it isn'a PTA for Marie while she is in the process of building and testing her machine, if her evidence for 2 is simply that 1 is true. She can't learn that the machine is accurate by simply trusting its readings. That's true even if it is, in fact, reliably accurate. Jonathan Vogel (2000) has argued that this is a problem for many forms of reliabilism. Stewart Cohen (2002, 2005) has offered a generalisation of Vogel's argument that threatens normative externalism plus evidentialism, and we'll return to Cohen's argument later in this chapter. But for now we just need to note that this argument is not a PTA for Marie, using her new machine, while it might be for other agents. A historian of science a century after Marie, trying to retrospectively figure out how accurate Marie's innovative machine was, could use this argument in their inquiry.

So when we say that an argument is, or is not, a PTA, we mean to be talking about a particular, contextually supplied, agent, using something like the methods for learning the premises that they actually use. The phrase 'something like' is obviously rather vague, but the vagueness shouldn't worry us overly, as it won't compromise the discussion to come.

We have already seen some valid arguments that are not PTAs. The argument from the Eiffel Tower to London might not be a PTA for anyone in any possible

world. There is a radical version of the view that inference and implication must be kept separate which says that there are literally no valid PTAs. On this view, we never learn by following arguments from premises to conclusions, and thinking we do is a sign one has not properly appreciated the inference/implication distinction. I doubt this view is right. It is worth being sceptical about how often we use valid arguments in inference, but do seem to be some cases where we do. This schema, for instance, seems to be one we can easily use.

1. a_1 is the most recent F, and it is G.
2. a_2 is the second most recent F, and it is G.
3. a_3 is the third most recent F, and it is G.
4. So the last three Fs are G.

For a concrete instance of this, let F be *President of the USA*, G be *is left-handed*, and the a_i be Barack Obama, George W. Bush and Bill Clinton, and imagine someone considering the argument in 2009. More generally, consider cases where G is a coincidental property of the last 3 (or more) Fs, and we see that the last few Fs have this property by simply working through the cases. The result is a conclusion that we learn simply by remembering the premises, and then doing a very simple deduction. So there are some PTAs, even if not every valid argument is a PTA.

The clearest example of a valid argument that is not a PTA, for any agent, is *A, therefore A*. By definition, a PTA is one where the agent could first learn the premise, and then, in virtue of that, later come to learn the conclusion. But one cannot first learn the premise of *A, therefore A*, and later come to believe the conclusion. For similar reasons, it will be rare that *A and B, therefore A*, could be a PTA for an agent, though perhaps there are some possible instances of this schema, and some possible agents, for whom this is a PTA.

Why isn't the argument about radioactivity a PTA for Marie? In some sense, we might say that it is because it is circular. Marie can't use her new machine to learn that one of the premises is true, then use the argument to learn that the machine is reliable. And then, presumably, go on to use the fact that the machine is reliable to defend the second premise of the argument. Something looks to have gone wrong.

It is tempting now to generalise from Marie's case to the principle that no argument whose conclusion is that a particular method or tool is reliable, and whose premises were based on that method or tool, could be a PTA. But this is too quick.

Or at least, as I'll argue in the next section, those of us who are not sceptics should think it is too quick.

9.3. Liberalism, Defeaters and Circles

In this section I discuss the following argument.

1. Normative externalism says that some arguments that exemplify defeater circularity are PTAs.
2. No argument that exemplifies defeater circularity is a PTA.
3. So, normative externalism is false.

I'm going to spend a bit of time setting up what defeater circularity is. But the basic idea behind premise 2 is that the principle suggested at the end of the previous section is true. And the idea behind premise 1 is that if we reject level-crossing and accept normative externalism, we end up committed to violations of that principle. I will mostly be concerned to argue against premise 2, though I'll note that there ways we could push back against premise 1 as well. The ideas of this section draw heavily on work by James Pryor (2004) and we'll start with an important distinction he draws.

Pryor distinguishes three different approaches epistemological theorists might take towards different epistemological methods. He offers labels for two of these approaches; I've added a label for the third that naturally extends his metaphor. In every case, we assume agent S used method M to get a belief in proposition p. And we'll say the proposition *M works* is the conjunction of every proposition of the form (*M represents that* q) $\rightarrow q$ for every salient q, where $\rightarrow$ is material implication. Then we have the following three views.[3]

Conservatism S gets a justified belief in p only if she antecedently has a justified belief that *M works*.

Liberalism S can in some circumstances get a justified belief in p without having an antecedently justified belief that *M works*, but in some other circumstances she can properly use M and not get a justified belief in p, because her prior evidence defeats the support that M provides for p.

[3] I'm modifying Pryor's views a bit to make these attitudes towards methods, rather than towards propositions; this makes everything a touch clearer I think. But I'm following Pryor, and the literature that has built up around his work, in focussing on justification rather than rationality. For reasons that I discussed in chapter 7, I would rather focus on rationality. I think the difference between the two concepts is not significant to this part of the discussion.

Radicalism As long as S uses *M* correctly, and *M* genuinely says that *p*, and *M* actually works, then no matter what evidence S has against *M works*, she gets a justified belief in *p*.

Whether conservatism, liberalism or radicalism is the most intuitive initial view will vary depending on which particular method we are considering.

Scientific advances naturally produce a lot of methods that we should treat conservatively. This is what we saw in the case of Marie and her machine; she couldn't learn things about how radioactive some things are until and unless she knew the machine worked. And that's true in general of new methods we develop. But it isn't true, isn't even intuitively true, of all methods.

Arguably we should be radicals about our most fundamental methods, such as introspection. A child doesn't antecedently need to know that introspection is reliable to come to have introspective knowledge that she's in pain. As long as introspection works, it isn't clear this is defeasible. If as the child grows up, she hears from some fancy philosophers that there is no such thing as pain, she might get some reasons to doubt that introspection works. But when she introspectively (and perhaps involuntarily) forms the belief that she's in pain, she knows she is in pain.

It is a little trickier to say which methods we should be liberals about. Pryor (2000) suggests that we should be liberals about perception. Many epistemologists, following C. A. J. Coady (1995) are liberals about testimony. They deny that we need antecedent reason to believe that a particular speaker is reliable, i.e., that that person's testimony work's before getting testimonial knowledge. But we shouldn't just believe everything we hear, so testimonial justification is defeasible.

Conservatism and radicalism are fairly well defined views. That is, the class of conservative views all share a strong family resemblance to each other, as do the class of radical views. The main thing we need to say about distinguishing different types of conservatism is that some conservatives have supplementary views that greatly alter the effect of their conservatism. For instance, the Cartesian sceptic is a conservative about perception who denies that we can believe perception works without having perceptual beliefs. But some other philosophers are conservatives about perception who also believe that it is a priori that perception works. Those positions will be radically anti-sceptical. So conservatism may have rather different effects elsewhere in epistemology, depending on what

it is combined with. But the basic idea that one can use *M* iff one has prior justification for believing *M works* gets us a fairly well defined region of philosophical space, as does the view that one can use *M* under any circumstances at all.

In contrast to conservativism and radicalism, liberalism covers a wide variety of fairly disparate theories. The liberal essentially makes a negative claim, antecedent justification for believing that *M works* is not needed for getting a justified belief that *p*, and an existential claim, there is some way of blocking the support *M* provides to *p*. Different liberals may have very different views about when that existential claim is instantiated.

A conservative-leaning liberal thinks that there are a lot of ways to block the support that *M* provides to *p*. One way to be a conservative-leaning liberal is to say that whenever S has any reason to doubt that *M works*, the use of *M* does not justify belief in *p*. Pryor's own view on perception is that this kind of conservative-leaning liberalism is true about perception. If any kind of liberalism about testimony is correct, then presumably it is a very conservative-leaning liberalism, since it is easy to block the support that testimony that *p* provides to *p*.

A radical-leaning liberal thinks that there are very few ways to block the support that *M* provides to *p*, even if in principle there are some. One natural way to be a radical-leaning liberal is that the support is blocked only if S believes, or is rational in believing, that *M works* is false. An even more radical view says that the support is blocked only if S knows that *M works* is false. A fairly radical form of liberalism seems intuitively plausible for memory; we are entitled to trust memories unless we have good reason to doubt them. It's worth keeping these radical forms of liberalism in mind when thinking about whether pure radicalism is ever true.

Pryor also notes an interesting way in which arguments can seem to be circular. He doesn't give this a name, but we'll call it defeater circularity.[4]

Defeater Circularity An argument exemplifies defeater circularity iff evidence against the conclusion would (to at least some degree) undermine the jus-

[4] I'm assuming throughout this chapter that it makes sense to talk about defeaters for beliefs. I actually don't want to commit to that being true. But the assumption is safe nevertheless. Dialectically, the situation is this. I'm trying to respond to the best arguments I know of that normative externalism licences a problematic form of circular reasoning. If the whole ideology of defeaters is misguided, there isn't any danger that a defeater based argument will threaten work. But I'm not going to have the defence of normative externalism rest on that ideological claim.

> tification the agent has for the premises. This is Pryor's Type 4 dependence; see Pryor (2004, 359).

It is important that Pryor uses 'undermine' here rather than something more general, like 'defeat'. Any valid one premise argument will be such that evidence against the conclusion will rebut, at least to some degree, the justification for the premises. But it won't be necessary that this evidence undermines that justification. If one reasons *X is in Ann Arbor, so X is in Michigan*, then evidence against the conclusion will rebut whatever evidence one had that X is in Ann Arbor. But that might not undermine the support the premise provides to the conclusion, or that the evidence supplies to the premise. If one thought X was in Ann Arbor because a friend said that they just saw X, the counter-evidence need not impugn the friend's reliability in general. It might just mean the friend got this one wrong.

It is not preposterous to think that arguments which exemplify defeater circularity are defective in some way. Indeed, it is not preposterous to think that they are not PTAs. If the falsity of the conclusion would undermine the premises, then the premises rely, in some intuitive sense, on the conclusion being true. And that suggests the argument is circular. And circular arguments are not PTAs. Or at least so we might intuitively reason.

Pryor argues that some arguments which exemplify defeater circularity are, in the language being used here, PTAs. He gives two arguments for this conclusion. First, he offers direct examples of arguments that he says exemplify defeater circularity, but which could, it seems, be used to form justified beliefs in their conclusions. As he notes, however, the intuitive force of these examples is not strong. His second argument is that defeater circularity arguments suffer from some other vice, such as a dialectical vice, and we confuse this for their not being sources of justification.

Most forms of liberalism imply that there will be good arguments that exemplify defeater circularity. If liberalism about M is true, and S can sometimes observe that she is using M, then she should be able to make the following argument, which we'll call the M argument.

1. p.
2. M says that p.
3. So M got this one right.

By hypothesis, this could be a way that S comes to know that M is working. Since liberalism about M is true, she doesn't need to know that antecedently to using p

to get the first premise. But the conclusion is obviously entailed by the premises. So it looks like it could be learned by learning the premises an doing a little reasoning. A kind of liberalism that says that whenever S recognises which method she is using, that method is blocked from providing support, would not licence this reasoning. But that's a kind of liberalism that doesn't seem particularly plausible.

But the *M* argument does exemplify defeater circularity, at least if we're assuming a not-too-radical form of liberalism about *M*. If S got evidence against the conclusion, that would trigger the clause saying that evidence that *M* does not work blocks the support that the agent gets for *p* by using *M*. That is, in the presence of such evidence, the first premise would not be supported. So we have the conditions needed for defeater circularity. So if some not-too-radical form of liberalism is true, then some arguments that exemplify defeater circularity can generate knowledge, and are in that sense not viciously circular.

This is all relevant to us because it is plausible that defeater circularity is the kind of circularity that's at issue in debates over Independence. Return again to Chiyoko and Aspasia, and recall the reasoning Chiyoko does.

1. 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
2. Aspasia believes that 2+2=5, and 3+3=7, and 4+5=8, and 7+9=15, while I believe that 2+2=4, and 3+3=6, and 4+5=9, and 7+9=16.
3. So, she got those four sums wrong, and I got them right.
4. It is likely that I would get at least one of them wrong if I'd taken the drug, and unlikely that she would get all four wrong unless she'd taken the drug.
5. So, probably, I have not taken the drug, and she has.
6. So I should not modify my beliefs about arithmetic in light of what Aspasia says; she has taken a drug that makes her unreliable.

This violates Independence. Chiyoko believes that Aspasia is the unreliable one because she calculated some sums, and realises that Aspasia got them wrong. And she uses this to conclude that disagreement with Chiyoko should not move her.

But where has Chiyoko gone wrong? If, as the defender of Independence insists, she should not have ended up where she did, where was her first mistake? All parties agree that statements like premise 2 are usable in debates. And step 3 follows from steps 1 and 2, presumably in a way that Chiyoko can realise. Step 4 is true, and isn't anything she has any reason to doubt. Step 5 follows from 4 in a simple way, so Chiyoko can sensibly go from 4 to 5. And step 6 follows from 5 on

any plusible theory of disagreement. One shouldn't modify one's beliefs in light of disagreement with someone who has taken accuracy-destroying drugs.

So the problem must be with step 1. Now it isn't immediately obvious what is problematic about step 1. But perhaps we can see the problem if we think about things in terms of defeater circularity. The argument from step 1 to step 5 does, plausibly, exemplify defeater circularity. If Chiyoko had reason to believe that step 5 was false, she would have arguably have a defeater for step 1. So here we have an argument that the normative externalist thinks is a perfectly good argument, indeed a PTA, and a kind of circularity that it exemplifies. I suspect this is probably the best case for the claim that normative externalists are committed to a dubious kind of circularity.

There is a tricky dialectical point here. The normative externalist need not themselves agree that Chiyoko's argument exemplifies defeater circularity. After all, they think that Chiyoko can reason well about arithmetic even if she has misleading evidence that she has been drugged. But it would be good to not have to rely on this aspect of the theory in order to defend the theory. So let's just note that the objection does to some extent rely on a premise the normative externalist may wish to question, and move on.

My main reply to the objection is that exemplifying defeater circularity cannot, in general, prevent arguments being PTAs. And that's because there is a general argument that there must be at least some PTAs that exemplify defeater circularity. Here's the argument for that conclusion.

1. Liberalism is true about some method of forming beliefs or other, though we aren't necessarily in a position to know which method it is.
2. If liberalism is true about some method of forming beliefs or other, then some PTAs exemplify defeater circularity.
3. So, some PTAs exemplify defeater circularity.

I think this argument can be found in Pryor (2004), though he spends more time on arguing that particular exemplifications of defeater circularity are PTAs than directly defending the existential claim.

I've already argued for premise 2, in the discussion of liberalism. And the argument is valid. So the important thing is to argue for premise 1. The main argument here is a scepticism-avoidance argument. I'm going to make an argument very similar to one found in reecnt work by David Alexander (2011) and Matthias Steup (2013). They both argue, and I agree, that otherwise plausible anti-circularity principles lead to intolerably sceptical conclusions. My version

of this argument goes via Pryor's notions of conservatism, liberalism and radicalism.

Call someone an *extremist* if they are anti-liberal about all methods. One way to be an extremist is to be a global conservative. The Pyrrhonian sceptics we will meet soon are global conservatives, and that's why they reach such implausibly sceptical conclusions. But there are more extremists than that. Someone who thought that for any method, either radicalism or conservatism is true of that method is an extremist in my sense.

It actually isn't too hard to motivate extremism. I suspect many philosophers would find the following argument at least somewhat plausible.

1. For any method of forming beliefs, either it is a priori knowable that it works, or it is not.
2. We should be radicals about any method of forming beliefs such that it is a priori that the method works.
3. We should be conservatives about any method of forming beliefs such that it is not a priori that the method works.
4. So we should be extremists about all methods.

For what it's worth, I think both premises 2 and 3 are false. But starting with this connection to the a priori helps bring out the connection between the argument against extremism and what I've written elsewhere about Humean scepticism (Weatherson 2005, 2014b). The problem with extremism is that it implies external world scepticism, and we should not be external world sceptics.

Why think that extremism implies external world scepticism? One strong reason is the long-running failure of anyone to come up with a plausible extremist response to sceptical doubts. To my mind, there is only one such response that even seems remotely plausible. This is the view that says we should be radicals about inference to the best explanation and introspection, plus the premise that the best explanation of our introspected phenomenology is that the external world exists. This kind of approach is defended, though not exactly in these terms, by Bertrand B. Russell (1912/1997, ch. 2), Frank Jackson (1977), Jonathan Vogel (1990), Laurence BonJour (2003), and other internalists.

Perhaps you think this kind of view can be made to work; my hopes for this project are dim. Let's just note one problem, one boldly conceded by BonJour. Since most humans have not justified their use of perception, etc by inference to the best explanation, it follows that most people do not have (doxastically) justified beliefs. That's implausible on its face, and it's symptomatic of a deeper

problem. Figuring out, or even being sensitive to, the quality of different explanations of the way the world appears is cognitively downstream from the kind of simple engagement with the world that we get in perception. So it is impossible to use inference to the best explanation to justify our belief that perception is reliable, at least if conservatism about perception is correct, because we need perception to make plausible judgments about the quality of explanations.

If that's all correct, then liberalism must be true about some methods. And that implies exemplifying defeater circularity cannot always be a bad-making feature of arguments. So the fact that normative externalists are committed to the goodness of arguments that exemplify defeater circularity cannot be, on its own, an argument against normative externalism.

And there is even more that the normative externalist can say. Assume I'm wrong in the last few paragraphs, and actually extremism is correct. Then we have a further question to ask: Is global conservatism correct or not? If not, some kinds of radicalism are correct. And if some kinds of radicalism are correct, then a strong form of normative externalism is true, at least with respect to beliefs formed by some methods. That's because radicalism implies that certain belief-forming methods are immune to all kinds of defeat, including belief that they don't work, or evidence that they don't work, or even knowledge that they don't work. That's a very strong form of normative externalism! Now it's true that what we get here isn't normative externalism in general, because all we get here is that for some belief-forming methods, higher order evidence is irrelevant. That's consistent with higher order evidence mattering sometimes, in a way that normative externalists deny. But if the position I'm imagining here - that higher order evidence is relevant to beliefs formed by certain methods - is correct, then general objections to normative externalism, ones that are insensitive to the methods by which people form beliefs, must be wrong.

On the other hand, if radicalism is never true, and extremism is true, then global conservatism is true. And global conservatism is a very implausible doctrine. To see how implausible, it's worth working through some varieties of sceptical argument.

9.4. Pyrrhonian Scepticism and Normative Externalism

In the previous section I argued that the principle that no PTA exemplifies defeater circularity leads to external world scepticm. But perhaps that was under-

stating the case. Perhaps it really leads to Pyrrhonian scepticism, and Pyrrhhonian scepticism is a kind of reason scepticism. (The next few paragraphs draw on a discussion of scepticism by Peter Klein (2015).)

Pyrrhonian scepticism starts with reflection on the problem of the criterion. Any knowledge we get must be via some method or other. But, says the Pyrrhonian sceptic, we can't use a method to gain knowledge unless we antecedently know that it is a knowledge-producing method. And plausibly that implies knowing it is reliable, since methods that are unreliable do not produce knowledge. So the Pyrrhonian is a global conservative, in the terminology of the previous section. Now knowing that a method is reliable is a piece of knowledge. So to know anything, there is something we need to know before we can know anything. That's impossible, so we know nothing.

The problem of the criterion is potentially a very strong argument. After all, the conclusion of the last paragraph was not that we know nothing about the unobservable, or about the external world, or even about contingent matters. It is that we know nothing at all. That even extends to philosophical knowledge. So the problem of the criterion is naturally an argument for Pyrrhonian scepticism, the view that we cannot know anything, even the truth of philosophical claims like Pyrrhonian scepticism.

For much the same reason, the view looks so strong as to be self-defeating. You might think that by the lights of the Pyrrhonian sceptic, we can't even assert Pyrrhonian scepticism, since we can't know it to be true. That's too quick, since it assumes as a premise that *Only assert what you know* is a valid rule, and that's both false in academic contexts, and easily denied by the Pyrrhonian. But still, a view that says we can't know that we exist, we can't know that we are thinking, we can't know that $\neg(0 = 1)$, and so on is just absurd.

And worse still, it is an argument for an absurd conclusion with really only one key premise, namely global conservatism. Sometimes arguments can have absurd conclusions, but at least they present us with a challenge to identify where things have gone wrong. Not here! The mistake is obviously the global conservatism, since that's the only premise there is. I'm assuming here that we are reading 'method' so weakly that it is uncontroversial that any knowledge is gained by some method or other.

And so most epistemologists do indeed reject that premise. Reliabilists say that any reliable method can produce knowledge, whatever the user of that method knows about the method's reliability. Other philosophers might say that we can

use induction in advance of knowing that induction is reliable, and hence in advance of knowing it is knowledge-producing. Or perhaps we can, as Descartes suggests, use clear and distinct perception before we know it is reliable. One way or the other, the overwhelming majority of epistemologists reject global conservatism somewhere.[5]

If global conservatism is false, then either liberalism is true somewhere, or radicalism is true somewhere. And we have already seen that either of these conclusions would be very bad news for circularity based objections to normative externalism. They certainly suggest that the argument from defeater circularity against normative externalism fails. If liberalism is true somewhere, then some PTAs exemplify defeater circularity, contra premise 2 of the argument. And if radicalism is true somewhere, then it is possible to be a normative externalist without committing to the view that the problematic arguments exemplify defeater circularity, contra premise 1 of the argument.

9.5. Easy Knowledge

The normative externalist looks like they will be subject to what Stewart (Cohen 2002, 2005) calls "The Problem of Easy Knowledge". This might be a better way to cash out the intuition that normative externalism leads to problematic kinds of circular reasoning.

The problem of easy knowledge arises for any theory that says an agent can use a method to gain knowledge without knowing that it is knowledge-producing. Say M is one such method, and S one such agent. And assume, at least for now, that S can identify how and when she is using M. That is, when she forms a belief that p using M, she at least often knows that she is doing so. Say that she forms beliefs p_1, ..., p_n this way, and each of these beliefs amount to knowledge. Then she can reason as follows.

1. $p_1 \wedge ... \wedge p_n$
2. M said that $p_1 \wedge ... \wedge p_n$
3. So, M is fairly reliable.

[5]The regress argument I've given here requires that the conservative view be stated a little carefully. It matters that the conservative says that M only provides justification if the subject antecedently believes, with justification, that *M works*. A view that says that M provides justification as long as *M works* was antecedently justifiably believable is not conservative as I'm carving up the space of views.

What could be wrong with this argument? We've assumed that the agent knows premise 1 and premise 2, so as long as she can use whatever she knows in an argument, she is in a position to run the argument. The argument is not deductive, but it seems like a decent inductive argument. Perhaps it could fail if there were external defeaters, but we can assume there are no such defeaters in S's situation. And if the sample size strikes you as too small for the inductive inference, we can increase the size of n.

So given some weak assumptions, it looks like S can use this argument to gather inductive support for the claim that M is fairly reliable. That is to say, she can use M itself to gather inductive support for the claim that M is fairly reliable. And that has struck many philosophers as absurd. This is, in essence, is the Problem of Easy Knowledge. Here are a few quotes from Cohen setting out what he takes the Problem to be. (The 'evidentialist foundationalist' in these quotes is the theorist who thinks that an agent can gain knowledge by drawing appropriate conclusions from evidence in advance of knowing that evidence reliably correlates with the appropriate conclusion. This is a form of normative externalism, and it's at least arguable that if Cohen's arguments work against the evidentialist foundationalist, they will generalise to all forms of normative externalism.)

> For example, if I know the table is red on the basis of its looking red, then it follows by the closure principle that I can know that it's not the case that the table is white but illuminated by red lights. Presumably, I cannot know that it's not the case that the table is white but illuminated by red lights, on the basis of the table's looking red. So the evidentialist foundationalist will have to treat this case analogously to the global deception case: I can know the table is red on the basis of its looking red, and once I know the table is red, I can infer and come to know that it is not white but illuminated by red lights. But, it seems very implausible to say I could in this way come to know that I'm not seeing a white table illuminated by red lights. (Cohen 2002, 313)

> It's counterintuitive to say we could in this way know the falsity of even the *alternative* that the table is white but illuminated by red lights. Suppose my son wants to buy a red table for his room. We go in the store and I say, "That table is red. I'll buy it for you." Having inherited his father's obsessive personality, he worries, "Daddy, what if it's white with red lights shining on it?" I reply, "Don't worry–you see, it looks red, so it is red, so it's not white but illu-

> minated by red lights." Surely he should not be satisfied with this response. Moreover I don't think it would help to add, "Now I'm not claiming that there are no red lights shining on the table, all I'm claiming is that the table is not white with red lights shining on it". But if evidentialist foundationalism is correct, there is no basis for criticizing the reasoning. (Cohen 2002, 314)

> Imagine again my 7 year old son asking me if my color-vision is reliable. I say, "Let's check it out." I set up a slide show in which the screen will change colors every few seconds. I observe, "The screen is red and I believe it's red. Got it right that time. Now it's blue and, look at that, I believe its blue. Two for two..." I trust that no one thinks that whereas I previously did not have any evidence for the reliability of my color vision, I am now actually acquiring evidence for the reliability of my color vision. But if Reliabilism were true, that's exactly what my situation would be. We can call this the problem of "easy evidence". (Cohen 2002, 317)

Cohen thinks that the lessons to draw from these cases is that we must distinguish between KR and PKR.

KR A potential knowledge source K can yield knowledge for S, only if S knows that K is reliable.

PKR A potential knowledge source K can yield knowledge for S, only if S has prior knowledge that K is reliable.

PKR is the problematic global conservatism. It leads to implausibly sceptical results. But, thinks Cohen, this is no argument against KR. Nothing in the discussion so far shows that there is anything absurd with a sweeping form of coherentism that says that S can to know simultaneously, and for the same reasons, all of the following propositions.

1. ¬(0=1).
2. I used a knowledge generating method to form the belief that ¬(0=1).
3. I used a knowledge generating method to form the belief that I used a knowledge generating method to form the belief that ¬(0=1).
4. I used a knowledge generating method to form the belief that I used a knowledge generating method to form the belief that I used a knowledge generating method to form the belief that ¬(0=1).

And so on. Cohen's opponents are the anti-coherentists who think it is possible to know ¬(0=1) prior to having this infinite chain of knowledge. Such anti-coherentists can, and do, disagree substantially about what exactly is required for one to know ¬(0=1). Let's start by considering just one opponent, a reliabilist who says that a method can produce basic knowledge if the following two conditions are met:

- The method is in fact reliable; and
- The agent has no reason to doubt that the method is reliable.

This is a somewhat simplified version of the reliabilism defended by Alvin Goldman (1986, 111–12), and similar in form (though not in its externalist commitments) to Pryor's dogmatism (Pryor 2000). And it is very much the kind of view that Cohen takes his arguments to be targeted against. He makes three observations about this kind of theory.

First, the theory allows for a fairly simple response to doubts grounded in sceptical possibilities. If something appears to be a red table, and so we come to know that it is is a table, we can simply deduce that we are not in a tableless room but deceived by an evil demon to think there is a table. This looks too quick, but as Cohen concedes, any response to scepticism will have some odd feature.

Second, the theory allows for a fairly simple response to more everyday doubts. This is the core of Cohen's objection to basic knowledge views. For instance, he notes that the kind of foundationalism that he considers would allow an agent to easily infer that they are not looking at a white table illuminated by red lights simply on the basis of the appearance of a red table. And this he thinks is absurd. This is the upshot of the first of the imagined conversations with his (then) 7 year-old son.

Third, the theory seems to allow for a fairly simple generation of grounds for an absurd inductive argument. Assume that the agent is living in a world where appearances do in fact reliably correlate with facts about the external world. So whenever something appears φ, the agent can know that it is φ, for any φ. So she can easily test the accuracy of her appearances just by looking. And the test will be passed every time, with flying colours! So she will have grounds for an inductive argument that appearances are an accurate guide to reality. This is the conclusion of the argument containing the second imagined conversion.

For now, let's assume that the intuitions about these cases are correct, and start with a question about the cases' significance. After bringing up intuitions about these few cases, Cohen makes some rather sweeping generalisations about the

impossibility of a plausible theory of basic knowledge. And that generalisation isn't supported by these cases.

Adding a defeasibility clause to foundationalism already avoids the worst of the problems. Cohen carefully distinguishes between inferences from everyday propositions to the falsity of outlandish sceptical claims, and inferences from everyday propositions (like *That's a red table*) to the falsity of other everyday-ish propositions (like *That's not a white table illuminated by red lights*). His reason for doing this is that it is the latter inferences that are especially implausible, since the necessity and difficulty of responding to the sceptic makes some otherwise counter-intuitive moves plausible. But once the defeasibility clause is in place, it isn't clear that the everyday cases are really problems. After all, if white tables illuminated by red lights are everyday occurrences, then the defeasibility clause will be triggered. And if they are not, we are back in the realm of sceptical doubts.

In other words, once the basic knowledge theorist adds a defeasibility clause, I don't think Cohen can avoid considering the kind of sceptical scenarios that he grants intuitions are unreliable about. It might be that the only things we can know by basic means are relatively simple anti-sceptical propositions, since we have reason to doubt everything else. Put another way, it's arguable that the unintuitiveness of Cohen's example is due to the fact that we have reason to doubt that the lighting is normal in a lot of examples. So my preferred foundationalist externalist will think it is not a case of basic knowledge. And anything they do think is basic knowledge won't be subject to these doubts.

To make this point more dramatically, consider the theorist (such as perhaps Descartes) who thinks that introspection is a form of basic knowledge. It is not unintuitive that we can see, by introspection, that introspection is reliable. We can introspect that p and introspect that we are introspecting that p, and so deduce that introspection worked on that occasion. At the very least, this isn't obviously wrong. For example, we mostly take our pain appearances to be reliable indicators of actually being in pain. They may or may not be reliable indicators of bodily damage, but they are reliable indicators of being in pain. We have no non-introspective evidence about this reliability. So we must, at some level, assume that introspection is good evidence that introspection is reliable.

Let's take stock. The big question is whether the Problem of Easy Knowledge helps us isolate a class of circular reasoning that is not acceptable. Cohen has demonstrated that several epistemological theories are committed to some reasoning that looks circular, like the reasoning involved in the imaginary conversa-

tions with his 7 year old son. Cohen himself takes those to be arguments against these epistemological theories, and by extension against a lot of circular reasoning. But it isn't clear that Cohen's arguments generalise as far as he intends; their intuitive force may turn on some special features about colour perception. So let's look more closely at the intuitions behind Cohen's examples.

9.6. What's Wrong with Easy Knowledge?

It's hard to put one's finger on just what is supposed to be wrong with easy knowledge. Cohen usually just relies on the intuitive implausibility of the methods he is discussing being knowledge producing. But it is hard to generalise from particular cases since intuitions about any given case might be based on particular features of that case. An explanation of the intuition would avoid that problem. So I'll go over a bunch of possible explanations of the intuitions Cohen is relying on, with the hope that once we know why these intuitions are true (when they are), we'll know how far they generalise.

Note that one simple explanation of intuitions in the cases Cohen gives is simply that radicalism, or even radical leaning liberalism, is wrong about colour perception. That would tell us something interesting about the epistemology of colour, but not something more general about knowledge and circular arguments. And it wouldn't be any kind of problem for the normative externalist, since the normative externalist as such has no commitments at all about the epistemology of colour perception.

The worry is that there is something more general behind Cohen's cases, something that will be general enough to raise a problem for normative externalism. I deny there is, but I don't think there is any way to back up this denial except to work through all the principles we might think are supported by Cohen's cases. So that's the game plan for this section. I'll set things up as a dialogue between an objector, who uses reasoning inspired by Cohen's cases to put forward views that are inconsistent with Change Evidentialism, and my responses to the objector. I'll generally leave off the arguments that the objector's positions are actually in conflict with Change Evidentialism, but mostly they are. There is one exception, where I make a fuss about this in the reply. The objector assumes that we are radicals, or at least radical leaning liberals, about perception in general. We could resist that, while holding on to Change Evidentialism, but I'd rather acquiesce in this assumption.

9.6.1. Sensitivity

> *Objection*:
> If you use perception to test perception, then you'll come to believe perception is accurate whether it is or not. So if it weren't accurate, you would still believe it is. So your belief that it is accurate will be insensitive, in the sense of Nozick (1981). And insensitive beliefs cannot constitute knowledge.

The obvious reply to this is that the last sentence is false. As has been argued at great length, e.g. in Williamson (2000, ch. 7), sensitivity is not a constraint on knowledge. We can even see this by considering other cases of testing.

Assume a scientist is trying to figure out whether Acme machines are accurate at testing concrete density. She has ten Acme machines in her lab, and proceeds to test each of them in turn by the standard methods. That is, she gets various samples of concrete of known density, and gets the machine being tested to report on its density. For each of the first nine machines, she finds that it is surprisingly accurate, getting the correct answer under a very wide variety of testing conditions. She concludes that Acme is very good at making machines to measure concrete density, and that hence the tenth machine is accurate as well.

We'll return briefly to the question of whether this is a good way to test the tenth machine below. It seems that the scientist has good inductive grounds for knowing that the tenth machine is accurate. Yet the nearest world in which it is not accurate is one in which there were some slipups made in its manufacture, and so it is not accurate even though Acme is generally a good manufacturer. In that world, she'll still believe the tenth machine is accurate. So her belief in its accuracy is insensitive, although she knows it is accurate. So whatever is wrong with testing a machine (or a person) against their own outputs, if the problem is just that the resulting beliefs are insensitive, then that problem does not preclude knowing those outputs are accurate.

9.6.2. One-Sidedness

> *Objection*:
> If you use perception to test perception, then you can only come to one conclusion; namely that perception is accurate. Indeed, the test can't even give you any reason to believe that perception is inaccurate. But any test that can only come to one conclusion, and

> cannot give you a reason to believe the negation of that conclusion, cannot produce knowledge.

Again, the problem here is that the last step of the reasoning is mistaken. There are plenty of tests that can only produce knowledge in one direction only. Here are four such examples.

First example. The agent is an intuitionist, so she does not believe that instances of excluded middle are always true. She does, however, know that they can never be false. She is unsure whether Fa is decidable, so she does not believe $Fa \vee \neg Fa$. She observes a closely, and observes it is F. So she infers $Fa \vee \neg Fa$. Her test could not have given her a reason to believe $\neg(Fa \vee \neg Fa)$, but it does ground knowledge that $Fa \vee \neg Fa$.

Second example. The agent is trying to figure out which sentences are theorems of a particular modal logic she is investigating. She knows that the logic is not decidable, but she also knows that a particular proof-evaluator does not validate invalid proofs. She sets the evaluator to test whether random strings of characters are proofs. After running overnight, the proof-evaluator says that there is a proof of some particular sentence S_0 in the logic. The agent comes to know that S_0 is a theorem of the logic, even though the failure of the proof-evaluatory to output that S_0 has a proof would not have given her any reason to believe it is not a theorem.

Third example. Ada has a large box of Turing machines. She knows that each of the machines in the box has a name, and that its name is an English word. She also knows that when any machine halts, it says its name, and that it says nothing otherwise. She does not know, however, which machines are in the box, or how many machines are in the box. She listens for a while, and hears the words 'Scarlatina', 'Aforetime' and 'Overinhibit' come out of the box. She comes to believe, indeed know, that Scarlatina, Aforetime and Overinhibit are Turing machines that halt. Had those machines not halted, she would not have been in the right kind of causal contact with those machines to have singular thoughts about them, so she could not have believed that they are not halting machines. So listening for what words come out of the box is one-sided in the sense described above; for many propositions, it can deliver knowledge that p, but could not deliver knowledge that $\neg p$.

Fourth example. Kylie is a Red Sox fan in Australia in the pre-internet era. Her only access to game scores are from one-line score reports in the daily newspaper. She doesn't know how often the Red Sox play. She notices that some days there are 2 games reported, some days there is 1 game reported, and on many days

there are no games reported. She also knows that the paper's editor is also a Red Sox fan, and only prints the score when the Red Sox win. When she opens the newspaper and sees a report of a Red Sox win (i.e. a line score like "Red Sox 7, Royals 3") she comes to believe that the Red Sox won that game. But when she doesn't see a score, she has little reason to believe that the Red Sox lost any particular game. After all, she has little reason to believe that any particular game even exists, or was played, let alone that it was lost. So the newspaper gives her reasons to believe that the Red Sox win games, but never reason to believe that the Red Sox didn't win a particular game.

So we have four counterexamples to the principle that you can only know p if you use a test that could give you evidence that $\neg p$. The reader might notice that many of the examples involve cases from logic, or cases involving singular propositions. Both of those kinds of cases are difficult to model using orthodox Bayesian machinery. That's not a coincidence. There's a well known Bayesian argument in favour of the principle I'm objecting to, namely that getting evidence for p presupposes the possibility of getting evidence for $\neg p$. The argument turns on the fact that this is a valid argument, for any values of E, H, x you like.

1. $\Pr(H) < x$
2. $\Pr(E) > 0$
3. $\Pr(H \mid E) \geqslant x$
4. So, $\Pr(H \mid \neg E) < \Pr(H)$

Intuitively, we might read this as saying that if E raises the probability of H above any threshold x, then $\neg E$ would be evidence against H. I haven't discussed that objection here, because it's irrelevant. When dealing with foundational matters, like logical inference, Bayesian modelling is inappropriate. We can see that by noting that in any field where Bayesian modelling is appropriate, the objection currently being considered works. What's not so clear, in fact what is most likely false, is that we can model the above four examples in a Bayesian framework. Bayesianism just isn't that good at modelling logical uncertainty, or changes in which singular propositions are accessible to the agent. But that's what matters to these examples.

9.6.3. Generality

> *Objection*:
> Assume we can use perception to come to know on a particular occasion that perception is reliable. Since we can do this in arbitrary

> situations where perception is working, anyone whose perception is working can come to know, by induction on a number of successful cases, that their perception is generally reliable. And this is absurd.

I'm not sure that this really is absurd, but the cases already discussed should make it clear that it isn't a consequence of Change Evidentialism. It is easily possible to routinely get knowledge that a particular *F* is *G*, never get knowledge that any *F* is not *G*, and no way be in a position to infer, or even regard as probable, that all *F*s are *G*s.

For instance, if we let *F* be *is a Turing machine in the box Ada is holding*, and *G* be *halts*, then for any particular *F* Ada comes to know about, it is *G*. But it would be absurd for her to infer that every *F* is a *G*. Similarly, for any Red Sox game that Kylie comes to know about, the Red Sox win. But it would be absurd for her to come to believe on that basis that they win every game.

There's a general point here, namely that whenever we can only come to know about an *F* only if it is a *G*, then we are never in a position to infer inductively that every *F* is *G*, or even that most of them are. Since even the foundationalist externalist doesn't think we can come to know by perception that perception is not working on an occasion, this means we can never know, by simple induction on perceptual knowledge, that perception is generally reliable.

9.6.4. A Priority

> *Objection*:
> Assume it is possible to come to know that perception is reliable by using perception. Then before we even perceive anything, we can see in advance that this method will work. So we can see in advance that perception is reliable. That means we don't *come* to know that perception is reliable using perception, we could have known it all along. In other words, it is a priori knowable that perception is reliable. (This objection is related to an argument by Roger White (2006), though note his argument is directed against a slightly different target.)

This objection misstates the consequences of the view that perception provides evidence when it works. If perception is working, then we get evidence for this

every time we perceive something, and reflect on what we perceive. But if perception is not working well, we don't get any such evidence. The point is not merely that if perception is unreliable, then we can't possibly know that perception is unreliable since knowledge is factive. Rather, the point is that if perception is unreliable, then using perception doesn't give us any evidence at all about anything at all. So it doesn't give us evidence that perception is reliable. Since we don't know antecedently whether perception is reliable, we don't know if we'll get any evidence about its reliability prior to using perception, so we can't do the kind of a priori reasoning imagined by the objector.

This response relies heavily on an externalist treatment of evidence. A first order internalist is perhaps vulnerable to this kind of objection. As I've argued elsewhere (Weatherson 2005), first-order internalists have strong reasons to think we can know a priori that foundational methods are reliable. Some may think that this is a reductio of this first-order internalism. (I don't.) But the argument crucially relies on first-order internalism, not just on foundationalism.

9.6.5. Testing

> *Objection*:
> It's bad to test a belief forming method using that very method. The only way to learn that a method is working is to properly test it. So we can't learn that perception is reliable using perception.

This objection is, to me, the most interesting of the lot. It is interesting because the first premise, i.e. the first sentence in it, is true. Testing perception using perception is bad. What's surprising is that the second premise is false. The short version of my reply is that in testing, we aim for more than knowledge. In particular, we aim for sensitive knowledge. A test can be bad because it doesn't deliver sensitive knowledge. And that implies that a bad test can deliver knowledge, at least assuming that not all knowledge is sensitive knowledge. Defending these claims is the point of the next section.

9.6.6. Circularity

> *Objection*:
> Even if we haven't put our finger yet exactly on the problem, the reasoning involved in getting easy knowledge is in some way circular, and we should be suspicious of it.

By this stage of the chapter, it should be clear what's wrong with this objection. The hope was that we would find some way of making the anti-circularity intuition more precise by investigating easy knowledge. But all we've ended up with is the view that easy knowledge is bad because it is in some vague sense circular. If this is the intuition behind the Problem of Easy Knowledge, we're back in the territory of the 'whiff of circularity' objection.

9.6.7. Multiple Properties

> *Objection*:
> Let's say we grant that each of the six properties you mentioned so far is individually compatible with knowledge. That doesn't show that every combination of them is compatible with knowledge. In general, $\Diamond p$ and $\Diamond q$ don't entail $\Diamond(p \wedge q)$. So you haven't shown easy knowledge is possible.

I don't quite know what to think about this objection. It strikes me as completely wrong-headed. The 'no easy knowledge' intuition seems, to me at least, to rest on an overlapping set of plausible but ultimately mistaken judgments about the relationship between knowledge, evidence and rationality/justifiability. I've argued that any possible reason one could have to support the intuition that easy knowledge is not knowledge is false, or not strong enough to support that conclusion. Could it be that the reasons work collectively when they don't work singularly? It's logically possible, but I don't see any reason at all to suspect it is true.

In short, there isn't any one reason to believe that the intuitions behind the most general form Problem of Easy Knowledge are correct. It could be that no one of them is correct, yet the intuitions are right because of some combination, or because of some extra factor. But at this stage, the best thing to do is to treat the intuitions as suspect. That means they can't form the basis for any objection to normative externalism, or any other theory.

9.7. Coda: Testing

In response to the 'testing' argument for the intuition that easy knowledge is no knowledge at all, I suggested that we should distinguish between a test being in general good and a test being the kind of thing which can ground knowledge. I

think that's true because tests also aim at sensitive belief. A test can fail in this aim, but still produce knowledge, because sensitivity isn't necessary for knowledge. Here's a simplified version of a real-life situation that makes that position somewhat intuitive.

> Inspection
> In a certain state, the inspection of scales used by food vendors has two components. Every two years, the scales are inspected by an official and a certificate of accuracy issued. On top of that, there are random inspections, where each day an inspector must inspect a vendor whose biennial inspection is not yet due. Today one inspector, call her Ins, has to inspect a store run by a shopkeeper called Sho. It turns out Sho's store was inspected just last week, and passed with flying colours. Since Sho has a good reputation as an honest shopkeeper, Ins knows that his scales will be working correctly.

Ins turns up and before she does her inspection watches several people ordering caviar, which in Sho's shop goes for $1000 per kilogram. The first customer's purchase gets weighed, and it comes to 242g, so she hands over $242. The second customer's purchase gets weighed, and it comes to 317g, so she hands over $317. And this goes on for a while. Then Ins announces that she's there for the inspection. Sho is happy to let her inspect his scales, but one of the customers, call him Cus, wonders why it is necessary. "Look," he says, "you saw that the machine said my purchase weighed 78g, and we know it did weigh 78g since we know it's a good machine." At this point the customer points to the certificate authorising the machine that was issued just last week. "And that's been going on for a while. Now all you're going to do is put some weights on the scale and see that it gets the correct reading. But we've done that several times. So your work here is done."

There is something deeply wrong with Cus's conclusion, but it is surprisingly hard to see just where the argument fails. Let's lay out his argument a little more carefully.

1. The machine said my caviar weighed 78g, and we know this, since we could all see the display.
2. My caviar did weigh 78g, and we know this, since we all know the machine is working correctly.
3. So we know that the machine weighed my caviar correctly. (From 1, 2)

4. By similar reasoning we can show that the machine has weighed everyone's caviar correctly. (Generalising 3)
5. All we do in testing a machine is see that it weighs various weights correctly.
6. So just by watching the machine all morning we get just as much knowledge as we get from a test. (From 4, 5)
7. So there's no point in running Ins's tests. (From 6)

Cus's summary of how testing scales works is obviously a bit crude but we can imagine that the spot test Ins plans to do isn't actually any more demanding than what the scale has been put through while she's been standing there. So we'll let premise 5 pass. (If you'd prefer more realism in the testing methodology, at the cost of less realism in the purchasing pattern of customers, imagine that the purchases exactly follow the pattern of weights that a calibrator following the guidelines of the officially approved methods of calibration.) If 3 is true, it does seem 4 follows, since Cus can simply repeat his reasoning to get the relevant conclusions. And if 4 and 5 are true, then it does seem 6 follows. To finish up our survey of the uncontroversial steps in Cus's argument, it seems there isn't any serious dispute about step 1.

So the contentious steps are:

- Step 2 - we may deny that everyone gets knowledge of the caviar's weight from the machine.
- Step 3 - we may deny that the relevant closure principle that Cus is assuming here.
- Step 7 - we may deny that the aim of the test is (merely) to know that the machine is working.

One way to deny step 2 is to just be an inductive sceptic, and say that no one can know that the machine is working merely given that it worked, or at least appeared to work, last week. But that doesn't seem very promising. It seems that the customers do know, given that the testing regime is a good one, and that the machine was properly tested, that the machine is working. And the inspector has all of the evidence available to the customers, and is in an even better position to know that the testing regime is good, so as step 2 says, she gets knowledge of the caviar's weight from the machine.

In recent years there has been a flood of work by philosophers denying that what we know is closed under either single-premise closure, e.g., Dretske (2005), or multi-premise closure, e.g., Christensen (2005). But it is hard to see how that

kind of anti-closure view could help here. We aren't inferring some kind of heavyweight proposition like that there is an external world. And Dretske's kind of view is motivated by avoidance of that kind of inference. And Christensen's view is that knowledge of a conjunction might fail when the amount of risk involved in each conjunct is barely enough to sustain knowledge. But we can imagine that our knowledge of both 1 and 2 is far from the borderline.

A more plausible position is that the argument from 1 and 2 to 3 is not a PTA. But that just means that Ins, or Cus, can't get an initial warrant, or extra warrant, for believing the machine is working by going through this reasoning. And Cus doesn't claim that you can. His argument turns entirely on the thought that we already know that the machine is reliable. Given that background, the inference to 3 seems pretty uncontroversial.

That leaves step 7 as the only weak link. I want to conclude that Cus's inference here fails; even if Ins knows that the machine is working, it is still good for her to test it. But I imagine many people will think that if we've got this far, i.e., if we've agreed with Cus's argument up to step 6, then we must also agree with step 7. I'm going to offer two arguments against that, and claim that step 7 might fail, indeed does fail in the story I've told, even if what Cus says is true up through step 6.

First, even if Ins won't get extra knowledge through running the tests on this occasion, it is still true that this kind of randomised testing program is an epistemic good. We have more knowledge through having randomised checks of machines than we would get from just having biennial tests. So there is still a benefit to conducting the tests even in cases where the outcome is not in serious doubt. The benefit is simply that the program, which is a good program, is not compromised.[6]

We can compare this reason Ins has for running the tests to reasons we have for persisting in practices that will, in general, maximise welfare. Imagine a driver, called Dri, is stopped at a red light in a quiet part of town in the middle of the night. Dri can see that there is no other traffic around, and that there are no police or cameras who will fine her for running the red light. But it is wise to stay stopped at the light. The practice of always stopping at red lights is a better practice than any alternative practice that Dri could implement. I assume she, like most drivers, could not successfully implement the practice *Stay stopped at red lights unless you know no harm will come from running the light*. In reality,

[6]The arguments of the next few paragraphs are obviously close to the arguments in Hawthorne and Srinivasan (2013).

a driver who tries to occasionally slip through red lights will get careless, and one day run a serious risk of injury to themselves or others. The best practice is simply to stay stopped. So on this particular occasion Dri has a good reason to stay stopped at the red light: that's the only way to carry out a practice which it is good for her to continue.

Now Ins's interest is not primarily in welfare, it is in epistemic goods. She cares about those epistemic goods because they are related to welfare, but her primary interest is in epistemic goods. But we can make the same kind of point. There are epistemic practices which are optimal for us to follow given what we can plausibly do. And this kind of testing regime may be the best way to maximise our epistemic access to facts about scale reliability, even if on this occasion it doesn't lead to more knowledge. Indeed, it seems to me that this is quite a good testing regime, and it is a good thing, an epistemically good thing, for Ins to do her part in maintaining the practice of randomised testing that is part of the regime.

The second reason is more important. The aims of the test are, I claim, not exhausted by the aim of getting knowledge that the machine is working. We also want a sensitive belief that the machine is working. Indeed, we may want a sensitive belief that the machine has not stopped working since its last inspection. That would be an epistemic good. Our epistemic standing improves if our belief that the machine has not stopped working since its last inspection becomes sensitive to the facts. Before Ins runs the test, we know that the machine will work. If we didn't know that, we shouldn't be engaged in high-stakes transactions (like the caviar sales) that rely on the accuracy of the machine. But our belief that the machine will work is not sensitive to one not completely outlandish possibility, namely that the machine has recently stopped working. After the test, we are sensitive to that possibility.

This idea, that tests aim for sensitivity, is hardly a radical one. It is a very natural idea that good tests produce results that are correlated with the attribute being tested. And 'correlation' here is a counterfactual notion. For variables X and Y to correlate in the releavnt sense just means that if X had been different, then Y would have been different, and the ways Y would have been different had X been different are arranged in a systematic way. When we look at the actual tests endorsed in manuals on how to calibrate balances, producing this kind of correlation looks to be a central aim. If a machine weren't working, and it were run through these tests, the tests would issue a different outcome than if the machine were working. But 'testing' the machine by using its own readings cannot produce results that are correlated with the accuracy of the machine. If

the machine is perfectly accurate, the test will say it is perfectly accurate. If the machine is somewhat accurate, the test will say it is perfectly accurate. And if the machine is quite inaccurate, the test will say that it is perfectly accurate. The test Ins plans to run, as opposed to the 'test' that Cus suggests, is sensitive to the machine's accuracy. Since it's good to have sensitive beliefs, it is good for Ins to run her tests.

So I conclude that step 7 in Cus's argument fails. There are reasons, both in terms of the practice Ins is part of, and in terms of what epistemic goods she'll gain on this occasion by running the test, for Ins to test the machine. That's true even if she knows that the machine is working. The epistemic goods we get from running tests are not restricted to knowledge. That's why it is a bad idea to infer from the badness of testing our eyes, say, using our eyes that we cannot get knowledge that way. The aims of tests don't perfectly match up with the requirements of getting knowledge.

10. Akrasia

The normative externalist seems to be committed to the following possibility. An agent, we'll call her Aki, has been given excellent arguments in favour of a false sceptical thesis. For concreteness, we'll assume the scepticism in question is testimonial scepticism. Nothing turns on the particular choice of sceptical thesis. But something does turn on whether there can be excellent arguments for any false sceptical thesis, and we'll return to this assumption below. For now we'll assume that Aki is confident that one cannot get reasons to believe propositions on the basis of testimony. And she is rational to be confident in this; it's what her philosophical evidence supports. But, we'll also assume, testimonial scepticism is false.

Aki now learns the proposition that a long-time friend, who has not lied to her in the past, said that p. She has weak probabilistic reasons to have greater credence in $\neg p$ than p, but these are the kinds of background reasons that are routinely overturned by testimony. The details don't matter, but if it helps to make the case concrete, imagine that p is the proposition that the home team won last night's baseball game, when it was known in advance that the away team was stronger, and was favoured to win. Upsets happen all the time in baseball, so a friend's testimony that the home team won should be only mildly surprising, and cause one to believe that the home team won. Since in this case the friend's testimony was caused by the fact that the home team did indeed win, it is doubly true that one should believe the friend.

And this is what Aki does. Despite her philosophical leanings, she can't bring herself to not believe what her friend says. That she can't follow her own views in this way shouldn't be surprising. The ancient sceptical texts are filled with both arguments for scepticism, and techniques for putting one's sceptical conclusions into practice. It was never assumed that mere belief in a sceptical view would suffice for control over one's mental states (Morison 2014). Aki is just like the people that the ancient sceptics were writing for; people who believed their views but could not put them into practice.

And of course, it's a good thing Aki does not have her theoretical doubts govern her beliefs. She gets a well-confirmed, and true, belief by trusting the testimony. Does she get knowledge? That's a hard question, turning on whether one thinks that knowledge is incompatible with this kind of mistake by one's own lights. I'm going to set that aside, and just focus on the fact that she gets a well-supported true belief. I think, though this is controversial, she gets a rational belief. So Aki is an epistemological case of what Arpaly calls inadvertent virtue. She forms the right belief, for the right reasons, while thinking these are bad reasons.

Normative externalism, of the kind I prefer, says that Aki is doing as well as she can in the circumstances. She is believing what her evidence supports. She violates a level-crossing principle, but since I'm arguing against level-crossing principles, I don't take this to be a problem. Good for Aki, a paragon of rationality!

This take on Aki's situation strikes many philosophers as implausible. Some philosophers go so far as to say that Aki's situation is literally impossible; we cannot truly believe of Aki that she both believes p and believes that this is an irrational belief (Hurley 1989; Owens 2002; Adler 2002). Many others think that Aki is possible but irrational; rationality requires that Aki keep her first-order and higher-order beliefs coherent, so if she has this combination of beliefs, she is irrational (Hookway 2001; Ribeiro 2011; Smithies 2012; Greco 2014; Horowitz 2014; Titelbaum 2015; Littlejohn 2018).

So we get the following argument.

1. If normative externalism is true, then some akratic attitudes are rational.
2. No akratic attitudes are rational.
3. So, normative externalism is false.

The short version of my response is that there is no understanding of 'akratic' that makes this argument plausible. We have to have a fairly expansive understanding of what akrasia is for premise 1 to be true. And on that understanding, premise 2 is implausible.

Note I'm using 'attitude' in a fairly expansive sense here. If one believes p and believes that it is irrational to believe p in one's situation, I'll call that combination an akratic attitude. This is perhaps non-standard - maybe we should say that's a pair of attitudes that only become a single attitude if one forms the conjunctive belief that p is true and irrational to believe. But distinguishing belief in a conjunction and belief in each conjunct would be needlessly distracting in this context. Put in other terminology, the best version of premise 2 will be a

'wide-scope' principle, saying that it is irrational to both believe *p* and believe that this very belief is irrational or otherwise defective.

10.1. The Possibility of Akrasia

I'm going to mostly assume that it is at least possible to, as Aki does, hold a belief while believing that very belief is in some way improper. I've tacitly given the argument for that assumption already. It draws on a very similar argument by Brian Ribeiro (2011). In practice there is a gap between, on the one hand, coming to accept a sceptical argument and being motivated to adjust one's mental life around it, and making those adjustments effectively. The very existence of Pyrrhonian techniques for resisting belief in propositions that one's theory says one should not believe is evidence of this gap. Anyone who falls into that gap, like Aki, will be akratic.

Could it be said that Aki doesn't really believe that sceptical arguments work? As David Owens (2002) points out, we don't want to just rely on Aki's firm avowals that she endorses testimonial scepticism; it takes more than talk to form a belief. But if Aki says that she endorses the sceptical arguments, and she tries to convince others of them, and she, for example, carefully studies Sextus Empiricus for strategies for putting her testimonial scepticism into effect, it seems plausible that she really does believe in testimonial scepticism. And that's true even if she lacks whatever it would take to put this sceptical doubt into full practice.

Is Aki, so described, akratic? Owens says that she is not, because she does not freely and deliberately choose to believe that the home team won last night, against her better judgment. Most other authors say, or perhaps just assume, that epistemic akrasia does not require freely and deliberately choosing one's beliefs. I'm not going to take a stand on the substantive question here. If we're trying to find a plausible version of the anti-externalist argument, it is best to not use 'akrasia' the way Owens does. That's because given Owens's usage, premise 1 is clearly false. Normative externalism makes no commitments at all concerning what it is rational to freely and deliberately believe. So let's assume we're working with a notion of akrasia that is not so demanding, and in particular that 'akrasia' applies to all cases where an agent believes against their better judgment.

10.2. Three Level-Crossing Principles

But even that characterisation is unclear on a key point. Here are three formulations of anti-akrasia principles that you could read as precisifications of the idea.

- "No situation rationally permits any overall state containing both an attitude A and the belief that A is rationally forbidden in one's current situation." (Titelbaum 2015, 261)
- "It can never be rational to have high confidence in something like *P, but my evidence doesn't support P*." (Horowitz 2014, 718)
- "If we use *Cr* for an agent's credences and Pr for the credences that would be maximally rational for someone in that agent's epistemic situation [then] $Cr(A \mid \Pr(A) = n) = n$" (Christensen 2010b, 122)

Titelbaum calls the principle he puts forward the 'Akratic Principle'. I don't want to use that name because part of what we're discussing is whether it is the most helpful way to understand akrasia. So I'll just call it Titelbaum's principle. Horowitz calls her principle the 'Non-Akrasia Constraint'. For similar reasons, I'll instead call it Horowitz's principle. The principle Christensen puts forward is commonly called *Rational Reflection*, and I'll follow that usage.

Rational Reflection is, in practice, considerably stronger than Titelbaum's principle. Imagine that Aki is having doubts about her testimonial scepticism. She doesn't fully endorse it. But she is still pretty confident in it; her credence in testimonial scepticism is 0.9. And she thinks that if testimonial scepticism is right, then the rational credence in the proposition that the home team won last night is below one-half. But she still has a very high confidence that the home team won, while thinking this is most likely irrational. This is a violation of Rational Reflection, but not of Titelbaum's principle. After all, there is no attitude that Aki both has and believes that it is irrational to have.

That doesn't show that Rational Reflection is logically stronger than Titelbaum's principle. Maybe there are states that violate Titelbaum's principle but not Rational Reflection. Whether this is so turns out to turn on difficult questions about the relationship between credence and belief. I'm not going to get into those questions here, in part because I have rather idiosyncratic views on them. On almost all theories about that relationship, however, it is impossible to violate Titelbaum's principle without violating Rational Reflection. That's what I mean by saying that in practice, Rational Reflection is a stronger principle.

Whether Rational Reflection is also stronger than Horowitz's principle is a little less clear. At first glance, it seems like it must be. Imagine someone whose credences are given by the following table:

Proposition	Credence
p	0.7
The rational credence for me to have in p is 0.7	0.9
The rational credence for me to have in p is 0	0.1

Such an agent violates Rational Reflection. Rational Reflection implies that an agent's credence in a proposition equals their expectation of the rational credence in that proposition. And the agent's rational expectation of the rational credence in p is, from the last two lines of the table, 0.63. But on the face of it, it doesn't look like they violate Horowitz's principle. There is no proposition they are both confident in, and confident their evidence does not support. So it looks like Rational Reflection is stronger than Horowitz's principle too. But the arguments below concerning iterated cases may cause us to doubt whether that's ultimately the case.

My view is that all three of these principles are false. It's a little trickier to say exactly which of the principles are inconsistent with normative externalism, and so must be rejected by anyone who accepts normative externalism. The simplest thing to say here uses the framework developed at the end of Part I, concerning core and peripheral commitments of normative externalism.

It is a core commitment of normative externalism that Rational Reflection is false. Rational Reflection offers a bidirectional link between what it is rational to believe, and what one believes about what it is rational to believe. And, at least as I read the proponents of the principle, the direction of explanation goes (at least in part) from the subject's beliefs about what is rational to facts about what is rational.

Just what to say about Horowitz's principle and normative externalism is less clear, because we need to see exactly how it applies in some tricky cases to get a sense of its scope. We'll return to this below.

On the other hand, it is a relatively peripheral commitment that Titelbaum's principle is false. Titelbaum's principle is only a one way connection. And it is at least possible to endorse it while thinking the order of explanation goes in an externalist friendly way. One might think that if one believes A, it is irrational to

believe that it is irrational to believe A in part in virtue of having that very first order belief. So there is at least a version of Titelbaum's principle for which the answers to all of the questions posed at the end of Part I is "No", and that makes it an extremely peripheral violation.

We get this very externalist friendly version of Titelbaum's principle if we think that rational beliefs must be true, at least when the belief is about the normative. Why might we think that? One way to motivate that view is to start with the arguments given by Clayton Littlejohn (2012) that only true beliefs can be justified, and try to either reason from there to the conclusion that only true beliefs are rational, or to amend the arguments so as that conclusion falls out. But another way is to argue that there is something special to normative beliefs. While descriptive beliefs can be false and rational, normative beliefs cannot. That is the lesson Titelbaum draws from his principle (which remember he calls the 'Akratic Principle').

> Ultimately, we need a story that squares the Akratic Principle with standard principles about belief support and justification. How is the justificatory map arranged such that one is never all-things-considered justified in both an attitude A and the belief that A is rationally forbidden in one's current situation? The most obvious answer is that every agent possesses a priori, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires. Not only can this reflection justify her in believing those facts; the resulting justification is also empirically indefeasible. (Titelbaum 2015, 276)

But even if Titelbaum's principle were true, it wouldn't support a conclusion nearly that strong. The inference here is of the form: Agents can't rationally form false beliefs about a particular topic, so agents have a priori justification for all possible true beliefs about that topic. And there are all sorts of ways to block that. We could say that all rational beliefs are true, as noted. Or we could simply say that for this topic, the truth of a proposition is a reason to believe it that is always strong enough to defeat rational justification to fully believe its negation. There are a lot of spaces between the claim that a proposition has a priori justification that can never be overridden, and the claim that that proposition can never be rationally believed to be false.

The upshot is that there are two distinct ways out, for the externalist, from the challenge posed by akrasia. One could adopt an extremely externalist epistemology of normative beliefs, as Titelbaum does. That will accept that akrasia is irrational, but deny that the core commitments of externalism entail that akrasia may be rational. Or one could accept that some forms of akrasia, such as violations of Rational Reflection, are rationally possible, and deny they are problematic. I'm going to take this second path. That's in part because it gives us a stronger form of externalism and I want to show how a strong form of externalism may be defended. And it's in part because that's the path I think is correct. Let's turn, then, to reasons that have been given for thinking that all forms of epistemic akrasia are problematic.

10.3. Why Not Be Akratic

I'm going to briefly discuss a simple, but bad, argument for thinking that all akratic agents are irrational. I'll call this the Argument from the Ideal. I don't think anyone in the current literature endorses this argument, so it should be uncontroversial that it fails. Indeed, I suspect it is relatively uncontroversial why it fails. But working through the argument will be helpful for getting to our main task, discussing the Argument from Weirdness. This argument turns on the following premise.

Weirdness is Irrational Akratic agents will say or do weird things, and only irrational agents would say or do those weird things.

I think Weirdness is Irrational is false, but the following similar principle is true.

Weirdness is Non-Ideal Akratic agents will say or do weird things, and no ideal agent would say or do those weird things.

Different forms of the argument from weirdness will occupy the rest of the chapter, and in every case my reply will have this form. Akratic agents do some odd things, weird things even, but this is evidence of their not being ideal, not of their being irrational.

But let's start with the Argument from the Ideal. Imagine a perfect agent, who is all knowing and perfectly good. For convenience, call this agent God. God will never be akratic. That's because God only believes things that are strongly supported by His evidence, and only believes truths, so He believes (truthfully!)

that everything He believes is supported by His evidence. This suggests a simple argument.

1. God is not akratic.
2. Rational people will, so far as they can, replicate God's properties.
3. So rational people will not be akratic.

The problem is that premise 2 has any number of counterexamples. As well as not being akratic, God is *opinionated*. By this, I mean that for any p, God will either believe p or believe $\neg p$. (I'm assuming here that if God exists then a kind of realism is true.) Does it follow that all rational people are opinionated? No, of course not. I don't know what the weather is like where you, dear reader, are. In many cases, I don't even know who you are, or when you are reading. So far, we might think this is just a failure of omniscience. But it doesn't mean that rationality requires that I be opinionated about who you are, where you are, when you are, or what the weather is like there then. Indeed, rationality requires that I not be opinionated about these questions. And that's true even though I know if I were ideal, I would be opinionated.

The point is not just that premise 2 of the Argument from the Ideal is false. It's that once we have the distinction between what would be ideal, and what would be rational in non-ideal circumstances, we can see how a lot of other arguments fail too. So let's start working through some of the Arguments from Weirdness with this distinction in mind.

It is plausible that in Aki's situation, where she believes p and believes the evidence does not support it, that she should say *p, but my evidence does not support p*. And this kind of Moore-paradoxical utterance is absurd, say some philosophers (Smithies 2012; Greco 2014); it's not something a rational person could say. And it's certainly weird, and non-ideal. But we can see that it could be rational by working through some other non-ideal cases.

Bulan isn't sure who she is. She is highly confident that Bulan's evidence is E_B. This is rational, though not quite right. She knows that E_B is weak evidence for q, and that her evidence is E_A, and that E_A is good evidence for q, and that q is true. And that's all good, because all of those things are true. She says *q, but Bulan's evidence does not support q*. It's hard to see what's wrong with that claim, and indeed even opponents of epistemic akrasia should not say it is irrational. It's only the distinctively first-personal claim, the one that we get when Bulan thinks her attitude is mistaken under a first-personal mode of presentation, that is problematic. That's interesting in itself; the Argument from Weirdness seems to rely on

a view about the distinctiveness of first-personal thought and talk. So there is a potential line of defence for the normative externalist that denies the critic's assumption that first-personal belief is special (Cappelen and Dever 2014). But let's grant the assumption that first-person thought and talk is special, and see what other ways we can raise problems for the Argument.

Imagine that Bulan now learns who she is. Since she can't hold on to all of the claims that she is Bulan, that her evidence is E_A, and that Bulan's evidence is E_B, she drops the middle claim. She instead holds on to the first and third claim, and infers that her evidence is E_B. Since she knows that E_B is weak evidence for q, she now believes that her evidence for q is weak. But since the fact that she is Bulan is no evidence against q, she also holds onto her belief that q. So now she thinks that *q, but my evidence does not support q*. And this is meant to be problematic, at least according to some opponents of epistemic akrasia. But it isn't at all clear which step was mistaken. I think that proponents of the Argument from Weirdness have to say that at the last step, one of two things must happen. Either Bulan must not resolve the tension in her beliefs by dropping the belief that her evidence is E_A, or she must take the fact that she is Bulan to be a reason to lose her belief in q, although her identity is probabilistically independent of whether q is true. Neither option seems appealing, and it's striking that proponents of the argument are forced into it.

Let's go back to the question of just what Aki (or Bulan) should say given their beliefs. Even if epistemic akrasia is possible, it doesn't immediately follow that rational agents will make these weird utterances. If it is only appropriate to say things if one knows them, as Williamson (2000) argues, and one can only know something if one's evidence supports it, then it can never be appropriate to say *p, but my evidence does not support p*. If one knows one evidence does not support p, then by the factivity of knowledge, one's evidence does not support p, so one does not know p, so one should not assert it. On this view, Aki shouldn't say *My evidence does not support p*, even if that proposition is supported by her evidence.

We don't need anything as strong as the rule *Only say what you know* to make the argument of the last paragraph work. Assume that for descriptive claims, the rule is *Only say what your evidence supports*, and for normative claims the rule is *Only say what is true*. Then if p is descriptive, it won't be permissible for Aki to say *p, but my evidence does not support p*. She will be able to say this if p is itself a normative claim. But the evidence that her assertion would be absurd in such cases is weak; there seem to be cases where this is exactly the right thing for her to say (Maitra and Weatherson 2010).

Horowitz (2014) carefully designs her akratic principle so as to ensure the arguments for it can't be so easily deflected. Imagine that Aki is more careful to not commit to anything that might be false. So she says *I'm confident that p, and I'm confident my evidence does not support p*. It is not plausible to say that one should only be confident in a proposition, or should only announce one's confidence in that proposition, if one knows the proposition to be true. For every lottery ticket in a large, fair lottery, I'm confident it will lose, yet I can't know each ticket will lose. (Perhaps I can't know any ticket will lose.) Horowitz argues that even this qualified utterance of Aki's is defective.

Notably, she doesn't just argue for this on the basis of intuitions about how weird the assertion itself sounds. There is a good dialectical reason for her to reason this way. The anti-akratic thinks that it is wrong to both be confident in p and in the proposition that the evidence for p is not strong, no matter which proposition p is, and no matter what the agent's background. It's hard to see how getting intuitions going about a few token utterances could support a universal generalisation that sweeping. So Horowitz offers some more careful arguments, ones that have at least the potential to generalise in the needed way.

Horowitz argues that Aki should be in a position to conclude, on the basis of her evidence, that her evidence is misleading, and that she was lucky to become so confident in the truth. And this, Horowitz thinks, is wrong. One needs independent reason to think that one's evidence is misleading, so it's wrong for Aki to conclude that on the basis of this very evidence. But that last premise seems too strong. Sometimes parts of one's evidence can be sufficient ground for thinking one's overall evidence is misleading. That's indeed what happens in Aki's case. There is no one part of her evidence that is both grounds for something and (complete) grounds for thinking those very grounds are misleading. The internal relations between the different parts of her evidence provide all the independent support we need for a reasonable judgment that other parts are misleading.

Horowitz has another argument that Aki will be in an untenable position. Imagine she is offered a bet that wins a small amount if p is true, and loses a larger amount if it is false. Aki takes the bet, as she should given that she has excellent reason to believe p is true. But she is then asked why she is doing this, she'll say that she should not be doing it; she has no good reason to believe the bet will win. Is this, doing something while saying one should not be doing it, problematic? Once we've seen other cases of inadvertent virtue, we can see why the answer is no. Huck Finn should help Jim escape, and should say he's doing the wrong thing while doing so. Aki's predicament is no worse.

Recently, Clayton Littlejohn (2018) argued for an anti-akrasia view by suggesting that Aki would end up with a distinct kind of untenable attitude. He imagines a conversation between Aki and her epistemic conscience with the following punchline. (Note in Littlejohn's example, the first order evidence supports not believing in p, and the higher-order evidence supports belief in p. This is the reverse of the case I've started with, but that doesn't matter much. What matters is that the levels diverge, and Aki follows the first-order evidence.)

> EC: You agree that it's irrational for you not to believe p. You agree that it's rational for you to agree on this point. You acknowledge that you don't believe p. You just don't yet see that this calls for any sort of change.
> Aki: Right. (Littlejohn 2018, 12, reference to preprint)

And this last statement of Aki's is untenable, thinks Littlejohn. And I suspect he is right about that. But it doesn't matter, because that's not what Aki should be saying. She should say that there is a "call for change", and she should think that there is such a call. After all, she thinks that she is not following her evidence, and that one should in general follow one's evidence. At the very least, that seems like reason to stop and have a think about how one got into this situation, and see if there wasn't some big mistake made along the way.

If Aki doesn't stop and reflect on her odd situation, that would be somewhat strange behaviour. But even the normative externalist can say that she should stop and reflect. It's true that she she isn't doing anything wrong. But whether one should stop and reflect is not entirely a function of whether one is doing anything particularly wrong. If one's cognitions or activities (or the conjunction of these) resemble those of people who are making mistakes, one has a reason to be think through what one has done. Of course, if Aki were ideal, she wouldn't need to stop and reflect, since she would know she is responding optimally to being in a strange situation. But if she were ideal, i.e., if she were God, she wouldn't be in that situation in the first place.

So we still haven't seen anything that Aki should do or say, given normative externalism, that is weird in a way that is inconsistent with rationality. She should perhaps say one thing and do another, just like Huck Finn. And she should say that Aki's evidence doesn't support what she herself believes, just like Bulan (in the original case) should say that Bulan's evidence doesn't support what she herself believes. But Huck Finn, and Bulan, aren't problematic. And the attempts to get Aki to say weirder things so far haven't worked; they've got her making assertions that violate norms of assertion even by the externalist's lights.

10.4. Self-Awareness and Rational Reflection

In the previous section I argued that there was nothing distinctively weird about akratic agents. They say and do weird things that other non-ideal but rational agents do. In this section I'll continue the argument a little, with more focus on two particular principles. In particular, I'll argue for these two claims:

1. Cases where agents do not know exactly what their situation is generate counterexamples to Rational Reflection, and to Horowitz's principle.
2. There is no reason to believe that these principles hold in cases where agents do know what their evidence is, since there is no reason to think that violations of the principles are more problematic in cases where agents do know what their evidence is.

I'll start with two relatively plausible assumptions:

1. What attitudes it is rational for an agent to have depend on features of her situation that vary from agent to agent and time to time.
2. The features that are relevant in point 1 are not luminous; agents might possess them without knowing that they do.

My view is that the 'features' in assumption 1 are just the agent's evidence, but I'm not assuming that. I'm just assuming that what's rational depends on the circumstances.

Premise 2 follows from the anti-luminosity arguments introduced by Williamson (2000), and defended recently by Hawthorne and Magidor (2009, 2011) and by Srinivasan (2015a). I don't need the full blown anti-luminosity principle to complete the argument. All I need is that luminosity fails for some of the features that are relevant to rational belief. So if there are some luminous states, as I've argued elsewhere (Weatherson 2004), that won't matter unless all features relevant to rationality are luminous. And that's not particularly plausible.

Even if all rational agents know exactly what is rationally required in all possible situations, as Titelbaum argues they do, there will still be failures of Rational Reflection. That is because an agent need not know what situation they are actually in. It is possible for an agent to have perfect knowledge of the function from situations to the rational status of states in such a situation, and not know what is rational for them. If rather extreme rational states are only permissible in rare situations, and the agent is in such a rare situation, then Rational Reflection will fail.

The abstract possibility described in the previous sentence is realized in Williamson's case of the unmarked clock (Williamson 2011, 2014). I'll work through Horowitz's variant, her case of the unmarked dartboard, because it provides a useful platform for setting up Horowitz's criticisms of the example, and my reply.

A dart is thrown at a dartboard that is infinite in height and width. The dartboard has gridlines on it running up-down and left-right. Due to magnets in the dart and the board, we know in advance that it will land on the intersection of two gridlines. The agent, we'll call her Siiri, can almost, but not quite, make out where it lands, and she knows in advance this will be the case.

Say that the 'distance' between two grid points, $\langle x_1, y_1 \rangle$ and $\langle x_2, y_2 \rangle$ is $|x_1 - x_2| + |y_1 - y_2|$. This is not the straight-line distance between the points; it is the shortest path between them on gridlines. Siiri knows in advance that if the dart lands on $\langle x, y \rangle$, then she'll know it is on $\langle x, y \rangle$ or one of the four points distance 1 away from it. And she knows in that situation it will be rational to have equal credence that it is on each of those five points.

Assume the dart lands on $\langle 8, 3 \rangle$, and consider her credence in the proposition that it is on $\langle 7, 3 \rangle$, $\langle 8, 4 \rangle$, $\langle 9, 3 \rangle$ or $\langle 8, 2 \rangle$. Call that proposition p. After getting visual evidence of where the dart is, her credence in p should be 0.8. But she should have credence 0.8 in p iff the dart is on $\langle 8, 3 \rangle$, and credence 0.2 in p if the dart is on any of the other four squares she thinks it might be on. So given her situation, the expected rational credence in p is 0.32. So Rational Reflection fails, even though Siiri knows exactly the function from situations to rational credences.

Horowitz argues that this is a special case. She thinks that a restricted version of Rational Reflection can be crafted that is immune to such a counterexample. There is something odd about the example. We're interested in a proposition p that is in a very odd class. Consider all propositions of the form *the dart lands distance 1 from point* $\langle$x, y$\rangle$. Siiri knows in advance that she will be very confident in such a proposition iff it is false. And that is odd. Here is how Horowitz puts the point. (Note that I've adjusted the terminology slightly to match what's here, and what she calls 'akrasia' is being highly confident in *p, but my evidence doesn't support p*.)

> In Dartboard, however, the evidence is *not* truth-guiding, at least with respect to propositions like p. Instead, it is *falsity*-guiding. It supports high confidence in p when p is false—that is, when the dart landed at $\langle 8, 3 \rangle$. And it supports low confidence in p when p

> is true—that is, when the dart landed at ⟨7, 3⟩, ⟨8, 4⟩, ⟨9, 3⟩ or ⟨8, 2⟩. This is an unusual feature of Dartboard. And it is only because of this unusual feature that epistemic akrasia seems rational in Dartboard. You should think that you should have low confidence in *p* precisely *because* you should think *p* is probably true—and because your evidence is falsity-guiding with respect to *p*. Epistemic akrasia is rational precisely because we should take into account background expectations about whether the evidence is likely to be truth-guiding or falsity-guiding. (Horowitz 2014, 738, notation altered, emphasis in original)

Surprisingly, it isn't essential to the example that the evidence is falsity-guiding in Horowitz's sense. This feature of the case is a byproduct of its simplicity; more complicated cases don't have this feature.

Imagine instead that when the dart lands at a particular spot ⟨*x*, *y*⟩, all spots whose distance from ⟨*x*, *y*⟩ is 10 or less are open epistemic possibilities for Siiri. But they are not equal possibilities; her probability distribution is peaked at ⟨*x*, *y*⟩ itself. For any grid point distance *d* from ⟨*x*, *y*⟩, her posterior probability that it landed there is:

$$\frac{4^{10-d}}{2,912,692}$$

The denominator there is just what's needed to make the probabilities add to 1. The intuitive idea is for each step further away from the center we get, the probability of being in that particular cell falls by a factor of 4. Now assume again the dart lands on ⟨8, 3⟩, though of course Siiri does not know this, and let *q* be the proposition that the distance between the dart and ⟨8, 3⟩ is either 0 or 3.

The evidence is not falsity-guiding with respect to *q*. Given what we said about Siiri, then among the worlds that are epistemically possible for her, her credence in *q* would be higher if *q* were true than if it were false. More precisely, her credence in *q* would somewhere between 0.413 and 0.44 if she were in one of the worlds that made *q*, and at most 0.307 if she were in one of the worlds that made *q* false. (The calculations to confirm the facts I'll run through about the example are tedious, but trivial, to verify with a computer.) The evidence supports higher confidence in *q* when *q* is true than than when *q* is false. That's unlike the original example. But this case also generates violations of Rational Reflection. Siiri's

credence in q is about 0.4275, but her expectation of the rational credence in it is about 0.3127.

Now you might think that's not a huge difference. Perhaps this is a counterexample to Rational Reflection, but not to Horowitz's principle that it is irrational to be highly confident in a proposition while also being highly confident that one is irrational to be so confident. But if we iterate the example, we get a counterexample to that principle too.

Imagine Siiri starts off (rationally) certain that repeated throws at the board are independent. And imagine that the dart is removed after each throw, so she can't see that successive darts land at the same spot. And imagine that her ability to detect where it lands doesn't improve, indeed doesn't change, over repeated throws. Finally imagine (somewhat improbably!) that repeated throws keep landing on $\langle 8, 3\rangle$. Let r be the proposition that at least 35 percent of throws are either distance 0 or distance 3 from $\langle 8, 3\rangle$. As the number of throws increases, she should get more and more confident that is true, and get more and more confident that it is irrational to think that it is true. After 100 throws, for example, her credence in r should be over 0.95, but her expectation of the rational credence in r should be under 0.25. This kind of iteration of examples can be used to turn any dartboard-like counterexample to Rational Reflection into a counterexample to Horowitz's principle.

10.5. Akrasia and Odd Statements

So Horowitz's explanation of why cases like Siiri's are special, that they are cases where agents know evidence is not truth-conducive, doesn't work. And that raises doubts for any attempt to separate Aki's case from Siiri's.

A large part of the motivation for thinking Aki's state is irrational is that Aki says weird things, like *p is true, although my evidence supports p being false*. But Siiri says similar things, and they are the right things for Siiri to say. So the very fact that Aki says them can't show that her position is incoherent; she is, in this respect, just like the perfectly coherent (if unfortunate) Siiri.

Siiri might regard it as a lucky break that she has a true belief despite not following her evidence. Of course, Aki could feel the same way. She should think that the home team won, think that her evidence doesn't support this, and from those claims think it is lucky that she has a correct belief despite not following

the evidence. But Siiri will think something structurally similar. Horowitz argues that Siiri doesn't have to regard herself as implausibly lucky. In the original version of the case Siiri knows the evidence is not truth-conducive, so it isn't a lucky break that not following the evidence (as it seems) leads to truth. But in the revised case, Siiri has to think she's just as lucky as Aki. And if it is reasonable for Siiri to think she is lucky, it is also reasonably for Aki to think she is.

Let's take stock. Siiri's case shows that Rational Reflection fails, and that it can be rational to be confident in something while also being confident that one's evidence does not support this view. It does not show that it can be rational to be confident in a falsehood about what rationality itself requires, as opposed to what one's situation is. That is, one could be certain about all the truths about what rationality requires in each situation, and still end up like Siiri. Indeed, we assumed she was certain about all the truths about what rationality requires in each situation, and still got a strange result falling out. So Siiri's case does not directly tell against the most plausible version of Titelbaum's principle.

But the arguments for Titelbaum's principle (or anything like it) are all Arguments from Weirdness. And Siiri's case does undermine the force of those arguments. For she says a lot of weird things too, and they are the right thing to say. So the fact that violations of Titelbaum's principle will lead to people saying weird, akratic things is no reason to think that Titelbaum's principle is a requirement of rationality. In weird situations, rational people are weird. Ideal people aren't weird, but that's only because they know things about their situation that are hidden from normal, rational, people. Normative externalism does imply that rational people will be akratic, and be weird, and be non-ideal. But none of that is surprising; the kinds of weirdness and non-idealness we see are just what we should independently expect in rational, but non-ideal, people.

10.6. Desire as Belief (Reprise)

The dartboard example is relevant to more than debates over akrasia. It also helps illustrate a point I alluded to frequently in part one, without ever setting out in detail. Proponents of the idea that moral uncertainty matters to rational decision making seem to be committed to a kind of 'desire as belief' thesis. David (Lewis 1988, 1996a) raised some technical problems for such theories, and recently those problems have been expanded by J. S. Russell and Hawthorne (2016). I'm not going to add anything to the arguments they have offered. But I think it might be helpful to translate those arguments into the idioms that are

more familiar in the moral uncertainty debates, since participants in that debate have not always appreciated the significance of these formal results. The only philosopher I know who has connected the moral uncertainty debates with the desire as belief debates is Ittay Nissan-Rozen (2015), and he takes an externalist position on moral uncertainty. My focus will be on the argument Russell and Hawthorne give, because it would be too much of a digression to investigate whether the 'desire by necessity' response that Huw Price (1989) gives to Lewis's arguments is successful.

Let's assume that we want moral uncertainty to play an important role in decision making. We should be able to provide some kind of semantics for claims about moral uncertainty. In particular, we would like a semantics for claims of the form *A is better than B* that satisfies the following four constraints.

1. Claims like *A is better than B* should be the kind of thing that can be believed, and that one can have higher or lower credences in. So that claim should be associated with a set of worlds, or a set of n-tuples, where the first member of that tuple is a world. (The latter disjunction is relevant if one thinks, perhaps following Lewis (1979), that the objects of belief are something like centred worlds.)
2. These attitudes in moral 'propositions' (or whatever else is picked out by *A is better than B*) should be updated in the way that credal attitudes are usually updated. Ideally that would be by conditionalisation, or by some other update rule that can be given independent motivation.
3. The semantics should associate with *A is better than B* a set of worlds (or tuples or whatever) that at least roughly corresponds with what those words ordinarily mean in English.
4. The claim should be action guiding, so (perhaps barring exceptional circumstances) conditional on *A is better than B*, *A* should be more choiceworthy than *B*.

And it turns out to be incredibly hard to find a semantics that satisfies these four constraints. In fact, there are principled reasons to think that no such semantics is possible.

There is one technical complication that we need to address first. Whether *A* is better than *B* depends on one's evidence. So if *A* is that I get a (typical) lottery ticket, and *B* is that I get a penny, then *A* is better than *B*, from my perspective, iff I don't know that *A* is a losing ticket. It is far from trivial to represent claims about what one's evidence is in a semantic model. That's in part because facts about what one's evidence is are 'first-personal' facts that are tricky to represent

in standard models, and in part because what one's evidence is changes over time, and it's hard to represent changes over time in standard models.

Here's how I'll try to deal with, or at least sidestep, these problems. Instead of thinking of beliefs as attitudes to sets of worlds, we'll think of them as attitudes to world-evidence-morality triples: $\langle w, e, m \rangle$. And we'll assume that e determines (perhaps among many other things) a function from times to one's evidence at that time. Just how it does that, and just how it attitudes distributed over e are updated, will be left as a black-box. (See Titelbaum (2016) for an excellent survey of the options for how the self-locating parts of one's credal state might be updated.)

I'll assume m is just a number, perhaps subject to enough constraints that we don't end up in the paradoxes of unbounded utility[1]. And what we want is that the value of a proposition is the expected value of m given that the proposition is true. So A is better than B, given some evidence, just in case the expected value of m given A and that evidence is greater than the expected value of m given B and that evidence. But expected values change with evidence, and evidence changes with time, so this doesn't settle what m should be. It turns out that while there are a few ways one could go here, any choice ends up violating one of the four constraints I proposed.

Assume, first, that the evidence is highly malleable. I mean two things by that. One is that when we conditionalise on some proposition c, then c gets added to the evidence. The other is that the time in question (and remember that e is a function from times to evidence sets) is the time any relevant decision has to be made. This pair of assumptions has a very nice feature - it guarantees that the fourth constraint is met. (This turns out to be harder to do than you might think.) Conditional on *A is better than B*, thus interpreted, I should choose A over B, no matter what the other evidence is.

The problem with this assumption is that it violates the third constraint rather dramatically. The following example is a version of the objection that J. S. Russell and Hawthorne (2016, 315–16) make to the principle they call Comparative Value. Consider the following substitutions for A and B.

A1 I get a can of frosty ice-cold Foster's Lager in five minutes time.
B1 I get a poke in the eye with a burnt stick in five minutes time.

[1] I'm assuming here that the moral value of a world can be represented as a number. That's not particularly plausible, but without this assumption the internalist views I'm opposing are very hard to state or defend.

I think that A1 is better than B1. And I even think that conditional on them both being true, which I hope they aren't. But on this model, we can't have that. Because conditional on them both being true, the expected value of m conditional on either of them is the same as the expected value m simpliciter. So conditional on their both being true, it isn't true that A1 is better than B1.

This is already a violation of constraint 3. But as Russell and Hawthorne go on to point out, a lot of strange things start to follow if we don't want to violate constraint 2 as well. We just proved that conditional on A1 $\wedge$ B1, it must be false that A1 is better than B1. That is, conditional on A1 $\wedge$ B1, the probability of A1 is better than B1 must be 0. If the way to update on A1 $\wedge$ B1 is by conditionalisation, it follows that the current probability of the conjunction of A1, B1 and A1 is better than B1 must be 0. So conditional on A1 is better than B1, which is surely true, the conjunction of A1 and B1 must have probability 0. And that's true for any A, B such that right now it's known that A is better than B. This is all absurd. Now perhaps this isn't a violation of constraint 2, because I'm assuming here that update is by conditionalisation, and maybe there is a principled way to reject that in cases like this. In any case, this option for how to understand e fails constraint 3, so it must be wrong.

The way this option failed suggested a distinct move. What's true about A1 and B1 is not that given they are both true, A1 will make the world better than B1 will. After all, given they are both true, they won't make any (further) difference to the world. So perhaps when assessing A1 and B1 for value, we should look at their initial value, or their value given the (absolutely) prior probability.

The problem with this approach is that it doesn't allow learning. Assume we learn C, than if I get poked in the eye with a burnt stick in five minutes, then malaria will be cured. Then it would be false that A1 is better than B1, and indeed true that B1 is better than A1. (Although, owww!) So this approach also violates constraint 3. And, for the same reason, it violates constraint 4.

Maybe the approach is to rigidify. What it means to say that *A is better than B* is that given the actual evidence I currently have, A has a higher expected m value than B. This will handle the the Foster's/poke case fairly well. But it leads to other problems. The following is a simple variant of the Rembrant case J. S. Russell and Hawthorne (2016, 331) offer.

Imagine we're in the simpler of the dart cases. When a dart lands on $\langle x, y\rangle$, then each of the five possibilities that it is on that very spot, or that it is one spot up, down, left or right are equally likely. And the dart did in fact land on $\langle 8, 3\rangle$. At

the same time, two fair coins have been tossed, although the results of them are hidden. Now compare the following options:

A2 I get a Vegemite sandwich if the dart landed on ⟨8, 4⟩, ⟨8, 2⟩, ⟨7, 3⟩ or ⟨9, 3⟩, and nothing otherwise.

B2 I get a Vegemite sandwich if at least one of the coins landed heads, and nothing otherwise.

Right now A2 is better than B2. That's because given my evidence, A2 gets me a 0.8 chance of a Vegemite sandwich, and B2 gets me a 0.75 chance. (Assuming, as is completely obvious, that more Vegemite sandwiches are better than fewer.) But conditional on A2 is better than B2, I should prefer B2. That's because the only worlds where A2 is better than B2 are worlds where the dart landed on ⟨8, 3⟩. And in those worlds, I don't get a Vegemite sandwich from A2.

So this rigid interpretation of 'better' violates constraint 4: it makes betterness judgments not be action guiding. I prefer A2 to B2, but conditional on A2 being better than B2, I prefer B2. Personally, I think this is the best interpretation of 'better', but that's because I think our choices shouldn't be guided by our beliefs about, or our evidence about, what's better than what.

I haven't given a watertight proof here that there is no way to interpret 'better' in this kind of model, or any other kind of model, that satisfies the four constraints. But philosophers who think moral uncertainty matters for decision making haven't typically appreciated how hard it is to get a model that does satisfy these constraints. The 'desire as belief' results are fairly surprising, and when combined with anti-luminosity principles, they make it very hard to see how moral uncertainty could be relevant to decision making.

11. Screening and Regresses

Normative externalism in epistemology is false if agents should respond not just to their evidence, but to what they believe, or should believe, about what their evidence supports. Call that latter claim the higher order hypothesis. Over the last three chapters I've responded to arguments for the higher order hypothesis. I argued that the cases that apparently support the higher order hypothesis do not do so, when viewed in the context of a wider sweep of cases. I've argued against attempts to derive the higher order hypothesis from anti-circularity principles. And I've argued against attempts to derive it from enkratic principles. In this chapter I move on to giving reasons to disbelieve the higher order hypothesis. I argue that the higher order hypothesis is tied to a principle about screening, a principle I call Judgments Screen Evidence. And I argue that this principle, whose name I'll shorten to JSE, leads to intolerable regresses. The return to regress based arguments provides a stronger link than we've seen so far between this part of the book and the earlier part; the arguments of this chapter might seem very familiar to someone who has read chapter 4.

11.1. Screening

The idea of screening that's going to be central to this chapter comes into philosophy via Hans Reichenbach (1956). He was working on a quite different problem, namely when we should infer that two events have a common cause. He says that C screens off the positive correlation between B and A iff the following two conditions are met.

1. A and B are positive correlated, i.e., $\Pr(A \mid B) > \Pr(A)$.
2. Given C, A and B are probabilistically independent, i.e., $\Pr(A \mid B \wedge \mathrm{C}) = \Pr(A \mid C)$.

I'm interested in an evidential version of screening. If we understand evidential support probabilistically, then we could just copy over Reichenbach's definitions, with a little reinterpretation of the formalism. So rather than thinking of Pr $ in terms of objective processes, as Reichenbach was, think of it as an evidential

probability function. Then these two clauses will say that as things stand, B is evidence for A, but given C, B is no evidence for A. We can say all that without assuming any particular connection between probability and evidence, as follows.

> C screens off the evidential support that B provides to A iff:
>
> 1. B is evidence for A; and
> 2. $B \wedge C$ is exactly as good evidence for A as C is.

Both these clauses, as well as the statement that C screens off B from A, are made relative to an evidential background. I'll leave that as tacit in what follows. Here are a couple of examples, the second loosely based on facts, that illustrate the usefulness of this idea.

A woman is standing at a suburban train station waiting for a train into the city, and wondering whether she will be on time for her meeting. She knows that there is only one train line, with no usable sidings, between where she is and the city, so there isn't any chance of trains passing. She knows how long trains take to get to the city if everything is working, though she doesn't know if everything is indeed working. But she doesn't know how frequent the trains are. She gets a call from a friend saying that a train to the city is headed her way, and is about five miles away. That train would, she thinks, get her to the city in time if everything goes right. Just then she sees a train coming into the station. Let A be that she gets to the city on time, B that there is a train five miles away, and C that there is a train pulling into the station. Relative to her initial background, B is evidence for A. But given C, it is no evidence at all. That's because given C, what matters is whether this particular train makes it in on time, without breaking down or being held up for some reason. The later train can't pass her, so its presence isn't relevant to whether she makes it to the city on time.

Later, she is trying to work out whether a particular person X voted for the Democratic candidate or the Republican candidate at the last Presidential election. She knows that X is either from Alabama or Massachusetts, and voted, and she knows the distribution of voters in those two states are as follows. (The numbers in the boxes are percentages of voters, and GOP is shorthand for the Republican Party.)

Pro-Choice	Pro-Life	Pro-Choice	Pro-Life
Dem	Dem	GOP	GOP

|:—|:—:|:—:|:—:|:—:| | Alabama | 28 | 7 | 7 | 58 | | Massachusetts | 52 | 13 | 13 | 22 |

Learning which state X is from is strong evidence about how they voted, since 65% of Massachusetts voters voted Democratic, while only 35% of Alabama voters did. But if she had previously learned that X was pro-choice, then learning which state X is from would be of no evidential significance. That's because 80% of pro-choice voters in each state voted Democratic. So learning that X is a pro-choice resident of Massachusetts is of no more evidential significance than simply learning X is pro-choice.

There is something very interesting about this theoretical possibility. We can concede that something is usually evidentially significant even while denying it is significant on a particular occasion. This possibility is useful for solving a puzzle about judgment.

11.2. The Counting Problem

Suppose a rational agent has some evidence E that bears on a proposition p, and on that basis judges that p. Call the fact that the agent has made this judgment J, and assume the agent is self-aware enough to know that J is true, and that she is rational. Assume also that p is a rational thing to judge on the basis of E, though the agent does not necessarily know this. The fact that a rational person judges that p seems to support p. After all, if we found out that she is rational and judged that p, that would ceteris paribus be evidence for p. Now consider this slightly informal question: *How many pieces of evidence does the agent have that bear on p*? Three options present themselves.

1. Two - both J and E.
2. One - E subsumes whatever evidential force J has.
3. One - J subsumes whatever evidential force E has.

This suggests a trilemma. First, it seems J could be evidence for p. We could get reason to be more confident in p just by learning J. Second, it seems like double counting for the agent to take both E and J to be evidence. After all, she only formed the judgment because of E. Yet third, it seems wrong for her to simply ignore E, since by stipulation it is evidence, and it certainly seems to bear on whether p is true.

One way out of this is to adopt the thesis I'll call JSE, for Judgment Screens Evidence. This is the thesis that propositions about rational judgments by rational agents screen off the evidential significance of the underlying evidence behind those judgments. The simplest argument for JSE is that it lets us answer the question above while accommodating the idea behind all three sources of 'pressure'. The agent can treat J just like everyone else does, i.e., as some evidence for p, without double counting or ignoring E. She can do that because she treats E as screened off. And screened off evidence isn't double counted or ignored. That's a rather nice feature of JSE.

To be sure, it is a feature that JSE shares with a view we might call ESJ, or evidence screens judgments. That view says that the agent shouldn't take J to be extra evidence for p,since its evidential force is screened off by E. This view also allows for the agent to acknowledge that J has the same evidential force for her as it has for others, while also avoiding double counting. So we need some reason to prefer JSE to ESJ.

One reason is by thinking generally about reasoning that proceeds in steps. Assume E is evidence for p solely because it makes q more likely, and q in turn makes p more likely. So if we are investigating a crime that took place in an inland village in Cornwall, learning that a suspect had some sand in his clothes that is only found on Cornish beaches may be some evidence that he's guilty. That's because it establishes that the suspect was at least in the area, unlike some other suspects. But if we knew independently that the suspect had been in Cornwall, say because he owns a beach house there and is often seen by his neighbours, the presence of the sand is of no evidential significance. Perhaps the general lesson here is that later steps screen off earlier steps. If that's right, we would expect J to screen E, and not vice versa.

Another reason for preferring JSE to ESJ is that it alone supports a number of positions that epistemologists have found independently plausible. Indeed, it is arguable that JSE is something of a tacit premise in a number of arguments. In the next section we will look at three such arguments.

11.3. JSE in Epistemology

11.3.1. Egan and Elga on Self-Confidence

We'll start with some conclusions that Andy Egan and Adam Elga draw about self-confidence in their paper "I Can't Believe I'm Stupid". I suspect many of the conclusions they draw in that paper rely on JSE, but I'll focus just on the most prominent use of JSE in the paper.

> One of the authors of this paper has horrible navigational instincts. When this author—call him "AE"—has to make a close judgment call as to which of two roads to take, he tends to take the wrong road. If it were just AE's first instincts that were mistaken, this would be no handicap. Approaching an intersection, AE would simply check which way he is initially inclined to go, and then go the opposite way. Unfortunately, it is not merely AE's first instincts that go wrong: it is his all things considered judgments. As a result, his worse-than-chance navigational performance persists, despite his full awareness of it. For example, he tends to take the wrong road, even when he second-guesses himself by choosing against his initial inclinations.
>
> Now: AE faces an unfamiliar intersection. What should he believe about which turn is correct, given the anti-reliability of his all-things-considered judgments? Answer: AE should suspend judgment. For that is the only stable state of belief available to him, since any other state undermines itself. For example, if AE were at all confident that he should turn left, that confidence would itself be evidence that he should not turn left. In other words, AE should realize that, were he to form strong navigational opinions, those opinions would tend to be mistaken. Realizing this, he should refrain from forming strong navigational opinions (and should outsource his navigational decision-making to someone else whenever possible). (Egan and Elga 2005, 82–83)

I will argue that this reasoning goes through iff JSE is assumed. I'll argue for this by first showing how the reasoning could fail without JSE, and then showing how JSE could fix the argument.

Start with a slightly different case. Katell is trying to find out whether p, where this is something she knows little about. She asks ten people whether p is true,

each of them being someone she has good reason to believe is an expert. The experts have a chance to consult before talking to her, so each of them knows what the others will advise. Nine of them confidently assure her that p is true. The tenth is somewhat equivocal, but says that he suspects p is not true, although he cannot offer any reasons for this suspicion that the other nine have not considered. It seems plausible in such a case that she should, or at least may, accept the supermajority's verdict, and believe p.

Now vary the case. The first nine are experts, but the tenth is an anti-expert. He is wrong considerably more often than not. Again, the first nine confidently assert that p, but now the tenth says the same thing, i.e., p. This doesn't change Katell's epistemic situation. She has a lot of evidence for p, and a little evidence against it. The evidence against has changed; it is now the confident verdict of an anti-expert, rather than the equivocal anti-verdict of an expert, but this doesn't matter. So she still should, or at least may, believe p.

Now make one final variation. Katell is the tenth person consulted. She asks the first nine people, who of course all know each other's work, and they all say p. She knows that she has a tendency to make a wrong judgment in this type of situation – even when she has had a chance to consult with experts. Perhaps p is the proposition that the correct road is to the left, and she is AE, for example. It does require some amount of hubris to continue to be an anti-expert even once you know you are one, and the contra-indicating judgments are made in the presence of expert advice. But I don't think positing delusionally narcissistic agents makes the case unrealistic. After listening to the experts, she judges that p. This is some evidence that $\neg p$, since she is an anti-expert. But, as in the last two paragraphs, it doesn't seem that it must override all the other evidence she has. So, even if she knows that in general she is fairly anti-reliable on questions like p, she need not suspend judgment. Even if her judgment is some evidence that $\neg p$, it might not be strong enough to defeat her earlier evidence for p. On those (presumably rare) occasions where her judgment tracks the evidence, the evidence may be strong enough for me to keep it, even once she acknowledges she have made the judgment.

The previous paragraph assumed that JSE did not hold. It assumed that Katell could still rely on the nine experts, even once she had incorporated their testimony into a judgment. That's what JSE denies. According to JSE, the arguments of the previous paragraph rely on illicitly basing belief on screened-off evidence. That's bad. If JSE holds, then once Katell makes a judgment, it's all the evidence she has. Now assume JSE is true, and that Katell knows herself to be something of an anti-expert. Then any judgment she makes is fatally self-undermining, just

like Egan and Elga say. When she makes a judgment, she not only has evidence it is false, she has undefeated evidence it is false. So if Katell knows she is an anti-expert, she must suspend judgment. That's the conclusion Egan and Elga draw, and it seems to be the right conclusion iff JSE is true. So the argument here relies on JSE.

11.3.2. White on Permissiveness

Roger White (2005) argues that there cannot be a case where it could be epistemically rational, on evidence E, to believe p, and also rational, on the same evidence, to believe $\neg p$. One of the central arguments in that paper is an analogy between two cases.

> Random Belief
> S is given a pill which will lead to her forming a belief about p. There is a ½ chance it will lead to the true belief, and a ½ chance it will lead to the false belief. S takes the pill, forms the belief, a belief that p as it turns out, and then, on reflecting on how she formed the belief, maintains that belief.

> Competing Rationalities
> S is told, before she looks at E, that some rational people form the belief that p on the basis of E, and others form the belief that $\neg p$ on the basis of E. S then looks at E and, on that basis, forms the belief that p.

White claims that S is no better off in the second case than in the former. As he says,

> Supposing this is so, is there any advantage, from the point of view of pursuing the truth, in carefully weighing the evidence to draw a conclusion, rather than just taking a belief-inducing pill? Surely I have no better chance of forming a true belief either way. (White 2005, 448)

There are two ways to read the phrase "from the point of view of pursuing the truth". One of them leads to an implausible view about the role of rational reflection in inquiry. The other makes the argument rely on JSE. Take these in order.

First, assume White's narrator is only concerned about having a truthful opinion right now, and only having a truthful opinion on this very question. Given that, it will be true that the belief-inducing pill will do just as well as careful weighing of the evidence. But that's a very unusual set of interests to have, and it's not clear why we should take such a person to show us much of interest about the point of reflection. One generally good reason for weighing the evidence carefully is that it puts us in a better position to be able to process new evidence as it comes in. It isn't clear how White's narrator, who takes the belief-inducing pill, will be able to adjust to new evidence, since by hypothesis he doesn't have any sense of how well entrenched this belief should be, and how sensitive it should be to counterveiling evidence. This point is closely related to the explanation Socrates gives for the superiority of knowledge to mere true belief in *Meno* 97d-98a.

Another good reason for weighing evidence carefully is that we learn about other propositions through this process. Assume we're trying to figure out whether p, and there is some other proposition q, such that (a) we care about whether q is true, and (b) p is sometimes, but not always, good evidence for q. It is very common that at least some such proposition exists. Then figuring out why p is true, or at least why we should think it is true, will be relevant for q. So an agent who only cares about having at this very moment a true belief about this very proposition might be no better off engaging in rational reflection than taking White's belief-inducting pill, but such agents are far removed from the usual situation we find ourselves in, and not good guides to epistemological generalisation.

But note that with JSE we don't need to restrict attention to such narrowly-defined agents. Assume that JSE is true. Then after S evaluates E, she forms a judgment, and J is the proposition that she formed that judgment. Now it might be true that E itself is good evidence for p. (The target of White's critique says that E is also good evidence for $\neg p$, but that's not yet relevant.) But given JSE, that fact isn't relevant to S's current state. For her evidence is, in its entirety, J. And she knows that, as a rational agent, she could just as easily have formed some other judgment, in which case J would have been false. Indeed, she could have formed the opposite judgment. So J is no evidence at all, and she is just like the person who forms a random belief, contradicting the assumption that believing p could, in this case, be rational, and that believing $\neg p$ could be rational.

Without JSE, White's analogy breaks down. Forming a belief via a pill, and forming a belief on the basis of the evidence, are very different. That's true even if you know that other rational agents take the evidence to support a different conclusion. The random belief is incapable of being properly updated, or of supporting

the correct strands elsewhere in the web of belief.

If we care about getting at the truth in general, and not just about *p*, then White's analogy needs JSE to go through. And we should, and do, care about truth in general. So this argument against permissiveness needs JSE. There may be other arguments against permissiveness, so this isn't to say that White's conclusion requires JSE. But his argument does.

11.3.3. Disagreement and Priority

Here is Adam Elga's version of the Equal Weight View of peer disagreement, a theory we will discuss much more in chapter 12.

> Upon finding out that an advisor disagrees, your probability that you are right should equal your prior conditional probability that you would be right. Prior to what? Prior to your thinking through the disputed issue, and finding out what the advisor thinks of it. Conditional on what? On whatever you have learned about the circumstances of the disagreement. (Elga 2007, 490)

It is easy to see how JSE could help defend this view. First, focus on the role JSE can play in the clause about priority. Here is one kind of situation that Elga wants to rule out. S has some evidence *E* that she takes to be good evidence for *p*. She thinks T is an epistemic peer. She then learns that T, whose evidence is also *E*, has concluded $\neg p$. She decides, simply on that basis, that T must not be an epistemic peer, because T has got this case wrong. This decision violates the Equal Weight View, because it uses S's probability that T is a peer after thinking through the disputed issue, not prior to it, in deciding who is more likely to be right.

Now at first it might seem that S isn't doing anything wrong here. If she knows how to apply *E* properly, and can see that T is misapplying it, then she has good reason to think that T isn't really an epistemic peer after all. She may have thought previously that T was a peer, indeed she may have had good reason to think that. But she now has excellent evidence, gained from thinking through this very case, to think that T is not a peer and so not worthy of deference.

Since Elga thinks that there is something wrong with this line of reasoning, there must be some way to block it. The best option for blocking it comes from ruling that *E* is not available evidence for S once she is using *J* as a judgment. That is, the best block available comes from JSE. Once we have JSE in place, we can

say what S does wrong. She is like the detective who says that we have lots of evidence that the suspect could have committed the crime–not only does he live in Cornwall, but he has Cornish sand in his clothes. To make the cases more analogous, we might imagine that there are detectives with competing theories about who could be guilty in this case. If we don't know who was even in Cornwall, then the evidence about the sand may favour one detective's theory over the other. If we do know that both suspects live in Cornwall, then the evidence about the sand isn't much help to either.

So JSE supports Elga's strong version of the Equal Weight View, which bars agents from using the dispute at issue as evidence concerning the peerhood of another. And if JSE is not true, then there is a kind of simple and natural reasoning which undermines Elga's Equal Weight View. So Elga's version of the Equal Weight View requires JSE.

11.4. JSE and Higher Order Evidence

As noted above, JSE can also support the higher order hypothesis. The idea is reasonably simple. Assume that an agent gets evidence that is in fact good evidence for p, concludes p on that basis, but also has reason to think they are in a suboptimal epistemic environment. The believer in higher-order evidence thinks the agent should then lower their confidence in p. But why is that, when they already have excellent evidence for p, and the evidence about the environment doesn't seem to defeat that?

Let's make that last rhetorical question a little clearer. Danail tells Milica that p. Milica has a long relationship with Danail, and he has been a very reliable testifier over that time. And Milica has no reason to doubt that p. But then Milica learns she has taken a drug that makes most people very unreliable when it comes to processing evidence by testimony. Should this last evidence reduce her confidence in p, by somehow defeating the support that Danail's testimony provides? The evidence about the drug isn't a rebutting defeater; it provides no reason to think p is false. But nor is it the most natural kind of undercutting defeater. It provides no reason to think that Danail is an unreliable testifier. What it does is undercut any support that Milica's own judgment gives to p. But that only matters to what Milica should believe if that judgment is playing an important role in sustaining her belief. And that's where JSE comes in. Unless JSE is true, Milica has a completely sound reason to believe p, namely Danail's testimony. And that reason isn't defeated by the drug. If a third party believed p

because they knew that Milica believed it on testimonial grounds, then the drug would be an undercutting defeater to the third party's belief. But to make it a defeater to Milica's belief, we need to assume that Milica, like the third party, in some way bases her sustained belief on her judgment. If JSE is right, then in a good sense she does do that; her own judgment that p is her only unscreened evidence, and if the force of it is defeated, then she has no good reason to believe p. If JSE is wrong, it is harder to see the parallel between Milica and the third party.

I've sketched an argument that the higher order hypothesis not just could be supported by JSE, but would be undermined if JSE were false. And JSE is indeed false, as we'll now show. We'll return at the end of the chapter to whether this fact can be turned into an argument against the higher order hypothesis.

11.5. The Regress Objection

Ariella is trying to make a forecast for how well her hometown team, the Detroit Tigers, will do in the upcoming baseball season. Baseball teams play 162 games[1], and the Tigers look like being a relatively mediocre team. She knows that it is irrational to form any belief about precisely how many games the Tigers will win. But she thinks, correctly as it turns out, that it is reasonable to form a credal distribution over propositions of the form *The Tigers will win n games*, and have that distribution be roughly a normal distribution with a standard deviation of 5 games. The question is to work out what the most likely win total is, which will be both the mode and the mean of the distribution. For simplicity, we'll say that for her to *predict* that The Tigers will win n games is to set n to be this centre. (I don't mean to suggest this is the ordinary use of the English word 'predict'. The definition I'm using here is stipulative.)

Ariella works through the known facts about the Tigers and their opponents, and predicts that they will win 76 games. This is, as it turns out, exactly the right prediction to make. That isn't to say the Tigers will actually win 76 games - remember the point here is not to form outright beliefs. Rather, the appropriate credal distribution over propositions about the Tigers' win total, given Ariella's evidence, is centred on 76.

[1] In reality they sometimes play 1 or 2 more or less. It will simplify the exposition to assume it is known in advance in Ariella's world that they play 162 exactly, and that's what I will assume.

But Ariella knows something about herself. She knows that in general, when she settles on a prediction, it is 1 game too low when it comes to the Tigers. If someone else knew nothing other than that Ariella had predicted the Tigers would win 76 games, and Ariella's track record, the rational thing for them to do would be to predict the Tigers will win 77 games. So Ariella has higher-order evidence that one might think will move her to change her prediction from 76 to 77.

Note carefully though what Ariella knows about herself. She knows that it is when she settles on a prediction that it is on average 1 game too low. If she decides that 76 wasn't a settled prediction, but 77 is, then she has exactly the same reason to raise her prediction to 78. And if she settles on that, she has a reason to raise her prediction to 79, and so on. Higher-order evidence is an issue because someone can have evidence that they make systematic mistakes in forming beliefs on the basis of evidence. But those systematic mistakes could also concern how they form beliefs on the basis of higher-order evidence. Indeed, they could be the same systematic mistakes in both cases. What should be done?

Let's start with three very bad ideas for Ariella. She should not simply follow the higher-order evidence where it leads, first raising her prediction to 76, then 77, then 78 and so on. After 87 steps, she will predict that the Tigers will win 163 games. Given that it is a 162 game season, this is not a good idea. Nor should she follow through as many steps of higher-order reasoning as she has the cognitive capacity to do. Assuming she has the ability to add 1 repeatedly, that will lead to the same flaw as above. And nor should she simply get out of the business of making predictions about baseball. (Compare Egan and Elga's comment that AE should simply stop making judgments about where to turn; a comment that was about one particular case of course, and not a general piece of advice.) Given what I've said so far about Ariella, she's really good at these kind of predictions. Having a small systematic error like this is not that much of a flaw, given how good she otherwise is.

There are three other strategies for dealing with higher-order evidence that are at least plausible. The first is the one I will defend. It is that Ariella should simply stick with her original prediction because it is the best prediction to make given her evidence. The second is that she should find some equilibrium point, where the higher-order evidence does not recommend a change of view. As stated, this view won't say anything about what Ariella should do, because there is no equilibrium position. But perhaps the view could be extended to say that she should follow the first-order evidence if there is no equilibrium, so it will also say that she should stick with her original prediction. The third option is that Ariella should follow one step of higher-order evidence, then stop with the prediction

that the Tigers will win 77 games. I'll argue for the first option by arguing against the other two.

Start with the idea that Ariella should, if possible, settle on an equilibrium. The idea is that we avoid the regress by saying that when possible, rational agents should be such that when they add the fact that they made that judgment to their evidence, the rational judgment to make given the new evidence has the same content as the original judgment. So if one is rational, and predicts that p, the rational prediction given that one has made the prediction that p is still p.

Note that this isn't as strong a requirement as it may first seem. The requirement is not that any time an agent makes a judgment (or prediction), rationality requires that they say on reflection that it is the correct judgment. Rather, the requirement is that when possible, rational agents make those judgments that, on reflection, they would reflectively endorse. We can think of this as a kind of ratifiability constraint on judgment, like the ratifiability constraint on decision making that Richard Jeffrey (1983) uses to handle Newcomb cases.

A judgment is ratifiable for agent S just in case the rational judgment for S to make conditional on her having made that judgment has the same strength and content as the original judgment. The regress is blocked by saying rational agents make ratifiable judgments when possible. If the agent does do that, there isn't much of a problem with the regress; once she gets to the first level, she has a stable view, even once she reflects on it.

This assumption, that only ratifiable judgments are rational, drives much of the argumentation in Egan and Elga's paper on self-confidence; it is a serious option. As the comparison to Jeffrey suggests, it has some historical pedigree. And though this would take much longer to show, it is probably the best way to make sense of the emphasis on equilibrium concepts in game theory. Nevertheless it is false. I'll first note one puzzling feature of the view, then one clearly false implication of the view.

The puzzling feature is that, as we have already seen, there need not be any ratifiable judgment to make. So the view will be somewhat incomplete. But maybe that isn't such a bad thing. We imagine the ratifiability theorist saying the following two things. (This isn't the only way to extend the ratifiability view, but I won't be objecting to this extension.)

1. It is important to make ratifiable judgments. Any judgment that is not ratifiable is not rational.

2. It is better, other things being equal, to have judgments that track the evidence.

This view will say that Ariella faces an epistemic dilemma. Anything she does will be to some extent irrational, since it will not be ratifiable. But the least bad option will be to predict that the Tigers will win 76 games, as she does. If you think epistemic dilemmas are impossible, you won't like this way of thinking about Ariella. But I don't think the arguments against epistemic dilemmas are particularly strong. If this was the worst thing to say about the ratifiability view then it would look like a reasonable view.

But it isn't the worst thing to say about the ratifiability view. The problem arises in cases where there is a ratifiable judgment. Change the case a little so Ariella doesn't tend to overpredict Tigers losses by 1 game; she tends to overpredict them by 1%. So if she predicts the Tigers will lose 86 games, an outsider going off that prediction and her track record wouldn't predict the Tigers will lose 85 games, they will predict the Tigers will lose 85.14 games. (Remember given the stipulated meaning of 'predict' we're using here, it can be perfectly sensible to predict that teams will win a fractional number of games. Indeed, there is no particular reason to think that the centre of the credal distribution over Tiger wins will fall on an integer. Remember also that there are 162 games in a season, so predicting 76 wins just is predicting 86 losses.)

Changing the expected error from a game to a percent doesn't seem like a big change at first blush. But now there is a ratifiable prediction for Ariella. It is that the Tigers will win 162 games, and lose 0. So if we think Ariella should make ratifiable predictions where possible, we should conclude that whatever her evidence about the Tigers hitting, pitching and fielding, she should predict they will win all 162 games in the season. This can't be right.

This kind of case proves that it isn't always rational to have ratifiable credences. It would take us too far afield to discuss this in detail, but it is interesting to think about the comparison between the kind of case I just discussed, and the objections to backwards induction reasoning in decision problems that have been made by Pettit and Sugden (1989), and by Stalnaker (1998). The backwards induction reasoning they criticise is a development of the idea that judgments should be ratifiable. And the clearest examples of when that idea fails are cases where there is a unique ratifiable judgment, and it is a judgment that first order considerations tell strongly against. The example of Ariella has, quite intentionally, a similar structure.

The other option for blocking the regress is to say that there is something special about the first revision. So if Ariella predicts that the Tigers will win 76 games, that screens her evidence about the Tigers' hitting, pitching and fielding. But if she changes her mind and predicts that they will win 77 games, on the basis of the higher order evidence, that doesn't screen her original prediction that they will win 76. So the regress doesn't even get started. This is structurally similar to a move that Adam Elga (2010) makes about disagreement. He argues that we should adjust our views about first-order matters in (partial) deference to our peers, but we shouldn't adjust our views about the right response to disagreement in the same way.

It's hard to see what could motivate such a position, either about disagreement or about screening. It's true that we need some kind of stopping point to avoid these regresses. But the most natural stopping point is before the first revision. Consider a toy example. It's common knowledge that there are two apples and two oranges in the basket, and no other fruit. (And that no apple is an orange.) Two people disagree about how many pieces of fruit there are in the basket. A thinks that there are four, B thinks that there are five, and both of them are equally confident. Two other people, C and D, disagree about what A and B should do in the face of this disagreement. All four people regard each other as peers. Let's say C's position is the correct one (whatever that is) and D's position is incorrect. Elga's position is that A should partially defer to B, but C should not defer to D. This is, intuitively, just back to front. A has evidence that immediately and obviously entails the correctness of her position. C is making a complicated judgment about a philosophical question where there are plausible and intricate arguments on each side. The position C is in is much more like the kind of case where experience suggests a measure of modesty and deference can lead us away from foolish errors. If anyone should be sticking to their guns here, it is A, not C.

The same thing happens when it comes to screening. Remove B from the example and instead assume that A has some evidence that (a) she has made some mistakes on simple sums in the past, but (b) tends to massively over-estimate the likelihood that she's made a mistake on any given puzzle. What should she do? One option, in my view the correct one, is that she should believe that there are four pieces of fruit in the basket, because that's what the evidence obviously entails. Another option is that she should be not very confident there are four pieces of fruit in the basket, because she makes mistakes on these kinds of sums. Yet another option is that she should be pretty confident (if not completely certain) that there are four pieces of fruit in the basket, because if she were not very

confident about this, this would just be a manifestation of her over-estimation of her tendency to err. The 'solution' to the regress we're considering here says that the second of these three reactions is the uniquely rational reaction. The idea behind the solution is that we should respond to the evidence provided by first-order judgments, and correct that judgment for our known biases, but that we shouldn't in turn correct for the flaws in our self-correcting routine. I don't see what could motivate such a position. Either we just rationally respond to the evidence, and in this case just believe there are four pieces of fruit in the basket, or we keep correcting for errors we make in any judgment and start a regress.

11.6. Laundering

In the definition of JSE, I said it was restricted to rational judgments. This was to avoid a simple counterexample to the view. (I'm indebted here to Vann McGee for pointing out the need for this.) Vieno is usually a pretty reliable judge, and he's not currently drunk or otherwise incapacitated. But he makes a mistake, as we all do sometimes, and forms the belief that p on the basis of massively insufficient evidence. This is rather irrational. Again, that's not to say that Vieno himself is irrational, but he does have a particular irrational view.

Now assume that JSE were true in an unrestricted form. Vieno is a generally reliable judge. That he believes p is, on its own, pretty good evidence for p. If the underlying evidence E is screened off, then arguably the overall evidence does suggest that p, so Vieno's belief does track his evidence after all. More generally, if unrestricted JSE is right, then it is impossible for someone who knows themselves to be generally reliable to have an irrational belief. So unrestricted JSE must be wrong.

But even if we restrict JSE to rational judgments, some problems remain. For one thing, we need some explanation of why such a restricted thesis should be true. That is, we need an explanation of why JSE should be extensionally adequate in just the cases where it agrees with ESJ. The normative externalist, who believes in ESJ, has a simple explantion for that. JSE is extensionally adequate when and only when it agrees with ESJ because ESJ is generally true. It isn't clear what could be a similarly good explanation of why a restricted version of JSE holds.

Thinking through cases like Vieno's can help motivate ESJ, and normative externalism more generally. There is something very strange about his case. On the one hand, the fact that a reliable person like Vieno believes that p should be

some evidence for *p*. On the other hand, if Vieno still knows why he believes that *p*, still knows that is that *E* is the relevant evidence on which the belief was based, then believing that *p* still seems irrational. And that's despite his knowing one important piece of evidence in favour of *p*, namely that he himself believes it.

It's important to distinguish the claims I've made in the last paragraph from what Gilbert G. Harman (1986) says about a slightly different case. Imagine that a month later, Vieno has forgotten the evidence that led to the belief that *p*, but nevertheless believes *p*. There are two interesting variants of this example. In one, *p* has been stored in preservative memory over that time. In the second, Vieno bases a new belief that *p* on the memory of believing *p* a month ago, plus his general reliability. If Vieno was under no obligation to retain the evidence for *p*, then it is plausible in the second case that the new belief that *p* is rational. And if the belief is rational in that case, maybe it is rational in the case where *p* was stored in preservative memory too.

We've already discussed memory in some detail. Here i want to distinguish the following two kinds of cases. In one, Vieno has an apparent memory that *p*. In the other, he has a clear memory that *E*, and irrationally infers *p* from that. In the second, Vieno's belief is irrational. But it is a mystery why this should be so, since he has this excellent evidence for *p*, from his own track record of success. ESJ explains this nicely, since that evidence is screened off. So the case of Vieno is both a problem for JSE, and a boon for ESJ. The case shows that JSE needs to be restricted, but it is hard to motivate any particular restriction. And ESJ offers a nice explanation of a puzzling fact, namely why Vieno's track record is not in this case evidence for *p*.

Now ESJ is a strongly externalist thesis. It says that facts about one's own judgment are not evidentially relevant to what judgment one makes, provided one has access to the evidence behind that judgment. And that suggests that the judgment should really just be judged on how it tracks the evidence, which is what the externalist says.

This point about laundering also offers a nice reply to a worry that I shouldn't have drawn a commitment to JSE from the passages I quoted above from Egan, Elga, White and Christensen. Perhaps they are only committed to a weaker thesis, something like that JSE is true when mistakes have been made, or when the agent has good reason to believe mistakes have been made. I didn't attribute such qualified theses to these epistemologists because the qualifications seem to make the theories worse. The qualified theories are still vulnerable to the

regress arguments that we drew out of the examples involving Ariella. And the point about laundering shows that JSE is most plausible when it is restricted to cases where mistakes have not been made.

Ariella's example doesn't just show that JSE is wrong. It gives us an extra reason to doubt the higher order hypothesis. If that hypothesis is true, then whatever prediction Ariella makes, she should raise her prediction as soon as she realises that she has made it. But that isn't plausible, since it leads from a reasonable starting point, a prediction of 76 wins, to an incoherent conclusion. So the higher order hypothesis is false, and the challenge it poses to normative externalism does not succeed.

11.7. Agents, States and Actions

With this discussion of regresses completed, we are in a position to evaluate an interesting alternative to my account of cases like Riika's. The alternative I'll discuss here says that if Riika does nothing in response to learning the higher order evidence, her resultant belief is perfectly acceptable, but this shows something bad about her. I'm going to first motivate such an alternative view, then suggest that the regresses we've discussed in this chapter pose a problem for it.

My account of Riika's example is somewhat conciliatory. I say it could be right for Riika to change her credences, depending on just how the case is filled in. But there is much to be said for the less conciliatory view that the only rational belief for Riika to have is the one she started with. After all, that's what her evidence supported, and she didn't get any counter-evidence. So how do we explain the intuition that it would be bad to not change her mind? By postulating a break between the evaluations of Riika's beliefs, on the one hand, and the evaluation of her actions, or of her, on the other.

It will help to have some slightly stipulative language available to discuss the cases. When agent S forms the belief that p, we can evaluate that belief, and the formation of it, in a number of distinct ways. First, we can ask whether the belief is well supported by her evidence. Let's say that the belief is *evident* if so, and not evident if not. Second, we can ask whether the belief is supported by the totality of her reasons to believe. Let's say that the belief is *rational* if so, and irrational if not. Third, we can ask whether an epistemically virtuous agent would have formed that belief. Let's say that the agent is *wise* if she is so virtuous at the time

the belief is formed, and unwise if she is not.[2] Fourth, and last for now, we can ask whether the practice the agent follows when forming the belief is one that she ought, all things considered, be following. Let's say her practice is *advisable* if so, and inadvisable if not.

What's crucial to evidentialism, as I conceive of it, is that the evident and the rational coincide. It does not commit itself on whether following the evidence is what wise agents do, or whether following the evidence is always advisable.

Just as we can make this four-way distinction among beliefs, we can make a similar four-way distinction among actions. An agent looks at the evidence in favour of different decisions, and then takes a decision. We'll assume, to simplify matters, that the agent has decent values in this process, so what's at issue is how the agent's doxastic system interacts with decisions to act. So we can describe actions as evident, rational, wise and advisable, with these terms having the same meanings as above.

With all these distinctions in mind, we can take another look at the cases that motivate higher-order theories. Consider, for instance, Adam Elga's example of the pilot who has evidence that it is possible they are suffering from hypoxia (Elga 2008). Is it obvious that it is irrational for them to believe that they have enough fuel for the trip, as their evidence supports?

Well, it does seem inadvisable for them to act as if they had enough fuel. But to get from premises about the the inadvisability of action to conclusions about the irrationality of belief requires a lot of steps. We could imagine reasoning as follows.

1. It is inadvisable to act as if one had enough fuel.
2. So, it is inadvisable to believe one has enough fuel.
3. So, it is unwise to believe one has enough fuel.
4. So, it is irrational to believe one has enough fuel.

Put this bluntly, every step seems questionable. There could be distinct norms of action and belief. There could be distinct norms of advice and evaluation. And there could be distinct norms that apply at the level of agents to those that apply at the level of individual beliefs. Let's look at these in order.

Once we see that there are a lot of distinct ways we can think the pilot goes wrong, it is wrong to insist that it is simply intuitive that the pilot has as irrational belief.

[2] In chapter 6 I noted that I'm using 'wise' for this kind of evaluation of agents, mostly following Maria (Lasonen-Aarnio 2010b, 2014a), though changing the terminology slightly.

The intuition is that something has gone wrong with the pilot; what in particular has gone wrong is a matter for theory. And perhaps what is being intuited is not anything at all about belief, but something about action. Perhaps it would be bad in some way to act on one's evidence, even if it would be rational to believe based on that evidence.

Allan Coates (2012) has developed a form of this response to the examples that motivate internalist accounts of higher-order evidence. It isn't just critics like Coates who have reacted in this way. Here is David Christensen making an argument that higher-order evidence matters to the rationality of belief.

> If you doubt that my confidence should be reduced, ask yourself whether I'd be reasonable in betting heavily on the correctness of my answer. Or consider the variant where my conclusion concerns the drug dosage for a critical patient, and ask yourself if it would be morally acceptable for me to write the prescription without getting someone else to corroborate my judgment. Insofar as I'm morally obliged to corroborate, it's because the information about my being drugged should lower my confidence in my conclusion. (Christensen 2010a, 195)

The thought for now is that the last line of this quote is simply false. There are all sorts of reasons it might be morally obligatory to corroborate even if the information about being drugged should not lower one's confidence. It's true that some forms of consequentialism about decision making will say that if confidence is not lowered, decisions should not change. But it is not at all compulsory to take a consequentialist attitude towards medical ethics. (Compare what Maria Lasonen-Aarnio (2014b, 430) says about rules governing the police.) And even if one is broadly consequentialist, Christensen's conclusion still does not straightforwardly follow.

We should take seriously the possibility that this is a case where agents should not change their credences, but should change how they act. Now that will be incoherent if you think that one should always maximise expected utility. But let's consider the possibility that this is a case where maximising expected utility is not the thing to do. It's a striking fact that the standard arguments for the propriety of maximising expected utility are almost always question-begging against the most interesting opponents (Maher 1997; Weatherson 1999). Imagine a theorist who says that the right thing to do is to maximise expected expected utility, and run your favorite argument for the properiety of maximising expected utility against them. In most cases you'll find at some stage you're just begging the

question against them. Consider, for instance, arguments based on representation theorems. These typically include as a premise that if the agent is choosing between two bets, and they have the same cost and same payoff, she should choose the bet that is the more probable winner. But this is just to assume that, in a special range of cases, she should maximise expected utility rather than expected expected utility, or anything else, and that assumption is, in this context, question-begging.

I don't mean this to be a serious argument against the view that we should maximise expected utility. Sometimes the best arguments for true positions are question-begging (Lewis 1982). And a whole chapter of this book defends the claim that we can learn from circular arguments. Indeed I believe for independent reasons that we should maximise expected utility. But I do think it is worth thinking about the fact that the relevant intuitions about higher-order evidence seem in the first place to be intuitions about actions, and require some substantive assumptions to generate conclusions about beliefs.

After all, if Riika should maximise expected expected utility, then she should order more tests, or get someone to confirm her diagnosis, before she acts. And that is true even if she actually has good evidence that the patient has dengue fever, as long as she lacks good evidence that she has good evidence. And perhaps that is what we are intuiting when we intuit that she should not act. The intuitions about the case, then are intuitions about action, but they don't imply anything about belief without a substantive theory of the action-belief connection (i.e., that one should maximise expected utility), and that theory lacks independent support.

This is a way of debunking the intuitions Christensen endorses about Riika's case. (And as noted many times, many other theorists have similar intuitions to Christensen's about similar cases.) As it stands, I don't accept this debunking story, because I accept the 'substantive theory of the action-belief connection', but this is a commitment that goes beyond normative externalism, and the rejection of level-crossing principles.

Let's assume that that bridge has been crossed though, and we have reason (either intuition or argument) to believe it would be inadvisable for Riika to believe her patient has dengue fever. What follows? Nothing much, unless we assume a very tight connection between assessments of agents and assessments of states, or between assessments of strategies and assessments of states. And there are very good reasons to separating these assessments. Maria Lasonen-Aarnio (2010b, 2014a) has argued for separating agent assessment from state as-

sessment, and argued that the standard intuition here involves conflating the two. And John Hawthorne and Amia Srinivasan have argued for separating assessment of states from assessment of strategies for coming to those states (Hawthorne and Srinivasan 2013).

Hawthorne and Srinivasan's argument is that these assessments come apart in general, so we should not be surprised if they come apart here. In general, it makes sense to distinguish between what someone should do in a particular circumstance, and what the person would do if they had instilled the habits that would be most effective in the long run. They give an example from sports. Their example involves tennis, but the idea generalises. Given the range of possible installable habits, it might be that the best habit to instil is one that will lead to expending valuable energy on occasionally chasing after lost causes. They are particularly interested in an epistemic limit on possible habits; the fact that we don't always know what we know means that we can't always react perfectly to our knowledge. But there are many possible limitations on possible habits due to our physical and cognitive limitations. And any one of these limitations will produce a gap between the optimal thing to do in a situation given one's knowledge or evidence, and what would be done if one had installed the optimal habits.

Now it may well be that the best habits we could have, given our cognitive nature, would involve second-guessing ourselves in cases like Riika's. Certainly if we think that our instincts involve some kind of 'optimism bias' (Sharot 2012), then it will be advisable to instil habits to counteract that bias. And it is very plausible that the fact that someone did something because they were acting on the best habit they could have is largely excusing. (I would say it is completely excusing, but I'm a little sceptical that there are complete excuses.) It seems plausible that our norms of advice are tied more closely to the idea of what the best habits are to instil, rather than to what is best to do in each situation, so the thing to advise someone to do just is what they would do if they had the best possible habits.

But all these facts should not obscure the fact that these are all second-best situations. Our cognitive and physical limitations mean that we sometimes cannot do what we should. That's why they are called limitations. So there are cases where the best thing to believe is what the evidence supports, but it is understandable and excusable to regard the matter as unsettled. And the grounds for the excuse are that agent has the optimal habit for situations like this. But as theorists we should not ignore the fact that optimising habits is a second-best solution. Best would be to believe and confidently act. And it would be best to believe and act not because this would be a lucky guess, but because one has sufficient reason to act.

So why didn't I just say all this in Chapters 7 through 9 rather than going through long detours about evidence and circularity? One reason is that we still need to explain the distinction between Riika's case and Raina's, and I'm not sure going via thoughts about advisability, wisdom or action will help with that. (This isn't a coy way of saying I think it won't; it's just that I haven't yet worked out a way to make it help.)

But a bigger reason is that we need to avoid the regresses. And the regresses suggest that policies like *Adjust one's credences to the higher-order evidence* are actually not optimal habits. That would be a bad habit for Ariella to adopt. And it would be bad to advise Ariella to adjust her credences to her higher-order evidence. There is no sensible way for her to comply with that advice, and it is bad to give advice that cannot be sensibly complied with. And it would be bad to let higher-order evidence guide Ariella's actions, since that would lead to betting on extreme results.

So I think that a broadly evidentialist approach is the best way to explain the cases. But it is worthwhile to note that there are good reasons to reject level-crossing principles about act or state evaluation, while accepting them about agent evaluation. And such approaches might end up saying more radical things about particular cases than I say in thoroughly rejecting level-crossing principles.

12. Disagreement

12.1. Introducing the Issues

So far in this book I have discussed issues about disagreement only insofar as they related to higher-order evidence. In this chapter I change tack, and consider questions about disagreement, and especially peer disagreement, in their own right.

Here is a schematic case of peer disagreement. Ankita and Bojan are peers in both of the following senses:

1. They know each other to be equally good at resolving a broad class of questions, of which the question of whether p is true is a representative member.
2. They know that they each have the same evidence that bears on p.

They then independently consider the question of whether p is true, and when they report back, it turns out they have different views. In one simple, if extreme, case, Ankita thinks it is true, while Bojan thinks it is false. What changes, if any, should they make to their judgments, once they know what the other thinks?

There are, as elsewhere in philosophy, slightly more actively defended answers to this question than there are philosophers working on the question. So we need to start not with a list of possible answers, but a taxonomy of them. The conciliationists say that Ankita and Bojan should, to a considerable extent, move their credences towards the other's. In the case where one believes p to be true and the other believes it to be false, they should move to both withholding judgment. The anti-conciliationists deny this; they say that at least in many cases, at least one of the two need not change their credences at all merely in light of the disagreement.

In theory, I'm an anti-conciliationist. In particular, I defend a view that I'll call the evidence aggregation theory of disagreement. When someone hears that a peer disagrees with them, that is defeasible evidence the peer has evidence that

they lack. Ideally, the hearer would work out exactly what that evidence is, add it to their stock of evidence, and react accordingly. That ideal is rarely, if ever, realised. In more realistic cases, the hearer assigns different probabilities to different hypotheses about what may have produced the disagreement. The typical case is that the peer has reacted differently to having different evidence to the hearer. And it is also typical that that's because the peer has evidence that the hearer lacks. So in (most) typical cases, the hearer will think there is evidence that they lack which supports the peer's view, and typically the rational reaction to learning that is to move one's views in the direction of the peer's view. But this credal movement is defeasible thrice over; sometimes the hearer knows the peer has reacted irrationally, sometimes the hearer knows the peer has strictly less evidence than they do, and (in very rare cases) the rational reaction to evidence of evidence for a rival view is not to move one's view towards the rival view.

In all cases, the guiding principle is that each party should be asking themselves, and each other, why does the other party have the views that they have?[1] If the most plausible answer is that the other party has information that is relevant to p, then one should adjust one's confidence in p to suit.

The evidence aggregation view of disagreement is really the conjunction of two separable views. The first claim is that the right theory of disagreement is a reason aggregation theory. That is, hearers should aggregate the reasons for belief they have with the reasons that their interlocutor[2] has for the conflicting belief. The second claim is that evidence, and evidence alone, provides reasons for belief. The focus of this chapter will be largely on the first claim, though the way I defend it will both presuppose the second claim, and indirectly provide some support for it.[3]

[1] Note that I'm assuming here that there is no doubt about what the other party's views are. In realistic cases there is usually doubt about this. But what we are interested in here is how one should rationally respond to learning that another person has views that differ from one's own. It is useful to think about this question separately from the question of whether one does really know, in a given situation, that the other person has different views. Obviously if someone reports having a very improbable view, we should not take that report at face value; they may be lying about what they believe. But as theorists we can still think about the question of how to react to others having different views, even radically different views.

[2] I'm going to talk about hearers and interlocutors for ease of exposition, but don't read much into this. It doesn't matter that the evidence the 'hearer' gets for the disagreement is testimonial. And as noted in the previous footnote, real life cases of testimony involve both questions about how probable it is that the interlocutor is sincere, and how one should react on the assumption that they are sincere. Our interest is solely in the latter question.

[3] There is a large epistemological literature on disagreement, but very little of it concerns what we should say in cases where non-evidential reasons for belief are allowed. We'll set those cases aside here, though I think the evidence aggregation theory handles them fairly smoothly.

One consequence of the evidence aggregation view is that a person who has got things right, i.e., responded correctly to the evidence, should not adjust their views if they know the other party has no evidence they lack. So it is anti-conciliationist about the extreme case we started with, where the parties know they have the same evidence. In practice, the evidence aggregation view disagrees with conciliationism less than you might expect. Given the expansive conception of evidence I have been defending, it is vanishingly rare that parties know they have the same evidence. Usually, the rational response to a disagreement is not to give high credence to the proposition that the other party has exactly the same evidence as one does. Instead, it is to give high credence to the proposition that there is evidence that one lacks, and that supports a view closer to that of one's interlocutor. This is what's right about conciliationism, but it is not what is usually defended by philosophical conciliationists.

The evidence aggregation view of disagreement that I'm promoting bears an obvious affinity to the justificationist view of disagreement that Jennifer Lackey (2010) defends. The main differences are really points of emphasis, not deep principle. Lackey describes her view as a way of taking the best features of each of conciliationism and anti-conciliationism; I'm interested in a version of the view that is clearly opposed to conciliationism. Relatedly, Lackey's explanations of some of the cases that motivate conciliationism are different to mine. But the similarities outweigh the differences, and I wanted to note her theory as the closest precursor to the theory I'll defend here.

Another big motivation for the this view of disagreement I'm defending comes from some remarks on testimony by Frank Jackson (1987). Jackson suggests that the primary role of testimony is evidence aggregation.

> Why should you ever accept what I say, unless you already did so before I spoke – in which case speech is a luxury? ... The answer cannot be that you are taking me to be sincere. ... Sincerity relates to whether you should infer prior agreement or disagreement in beliefs, not to whether posterior adjustment of belief is in order. The reason posterior adjustment in belief may be in order is that hearers (readers) sometimes have justified opinions about the evidence that lies behind speakers' (writers') assertions. You assert that P. I know enough about you, and your recent situation, to know (i) that you have evidence for P, for you would not otherwise have said it, and (ii) that your evidence is such that had I had it, I would have believed P. I borrow your evidence, so to speak. Typically, I won't know exactly what your evidence is. Perhaps you visited a factory

> and came back and said 'The factory is well run'. I don't know just what experiences you had there – just what you saw, heard, smelt and so on – but I know enough to know that had I had these experiences – whatever exactly they were – I too would have come to believe the factory well run. So I do. ...in this way an epistemological division of labour is achieved. Imagine the work (and invasion of privacy) involved if we all had to duplicate each other's evidence. Of course, I may not come to believe exactly what the speaker or writer believes. A friend returning from overseas may say to me of a certain country 'It is very well run'. I may know enough of my friend to know that experiences that would make him say that, are the kind that would make me say 'Dissent is suppressed'. In this case, I will borrow his evidence to arrive, not at what he believes, but at what I would have, had I had his experiences. (Jackson 1987, 92–93)

I agree with almost all of this, though I'm not going to issue a full defence of such an evidence aggregation account of testimony here. (Why 'almost'? Because it will be rather important later that we not be able to move as freely between sharing experiences and sharing evidence as Jackson does in the last line.) Rather, I'm just going to acknowledge my debt to Jackson's ideas, and move to disagreement.

I'm hardly the first person to start with broadly evidentialist intuitions and end up with anti-conciliationist conclusions about disagreement; you can see a similar trajectory in recent work by Maria Lasonen-Aarnio (2013, 2014a), and what I say here also owes a lot to her. But the details are different enough to justify a new variant on similar themes.

12.2. Two Concepts of Peerhood

My setup of the Ankita/Bojan case is ambiguous at a key point. I said that Ankita and Bojan are equally good at resolving questions like this. There are two natural ways to interpret this. We could read it as meaning that they are equally likely to come up with a rational verdict, or that their verdicts are equally reliable. David Christensen (2016) is very good on the importance of this distinction.

> The literature typically concentrates on people one has (independent of one's views on the disputed issue) good reason to take

> as *epistemic peers*–as rough equals along certain dimensions of epistemic evaluation. One such dimension concerns the evidence the other person has relevant to the disputed issue, and the other concerns how well she forms beliefs on the basis of her evidence. ...[W]e should notice that there are a couple of different ways of approaching the second dimension of evaluation–ways which are not always clearly separated. One focuses on the other person's equal likelihood of responding *rationally* to her evidence. On this reading, ... the disagreeing friend is what might be called a "rationality-peer" on the given issue: one whose opinion is equally likely to be rational. The second way of evaluating the other person's responses to evidence is in terms of her likelihood of responding to that evidence by forming *accurate* beliefs. On this reading, ... the disagreeing friend ... might be called an "accuracy-peer" on the given issue: one whose opinion on the disputed issue one expects to be as likely to be accurate as one's own. (Christensen 2016, 3)

Christensen cites Feldman (2007), Kelly (2005), Christensen (2007b) and Cohen (2013) as writers who understand peerhood in terms of rationality, and Elga (2007), White (2009), Enoch (2010), Kelly (2010), Lam (2011) and Levinstein (2013) as writers who understand it in terms of accuracy. He's not the first to notice these two possible understandings; the distinction plays a big role in work by Ben Levinstein (2013) and by Miriam Schoenfield (2014).

The rationality-based understanding is most relevant to the broader themes of this book. If peerhood is understood in terms of rationality, then the motivation to conciliate in light of peer disagreement is indirect. The conciliationist says that Ankita should do two things in light of Bojan's disagreement. First, she should use that disagreement as evidence that her initial view is irrational, then second, she should that fact as grounds for revising that first-order credence. Normative externalism disagrees with the second step. The fact that she has some higher-order evidence that she is irrational need not, on its own, be any reason to revise her first-order credence.

But many writers have noted that the first step of this sequence is dubious too. The most that Ankita gets from Bojan's disagreement is evidence that some view other than hers is rational. It does not follow that her view is irrational, unless we have make a background assumption that there is only one rational response to any given evidence. So it seems that the argument for conciliationism requires the thesis Roger White (2005) calls Uniqueness: that there is a single rational

response to evidence. Whether this seeming is really correct is actively debated: see Douven (2009), Kelly (2010), and Ballantyne and Coffman (2011, 2012) for interesting moves in the debate. I'm going to mostly not take a stance on this, since the arguments for conciliationism have other weaknesses.

It might seem that once Ankita views Bojan as an accuracy-peer, issues about higher-order evidence aren't relevant to determining whether she should conciliate in light of her disagreement. After all, in that case Ankita has two pieces of evidence; her own judgment and Bojan's. By hypothesis, each of them are equally accurate. So she should act as if she had two measuring devices, one which said that p was true, and the other that said it was false. And in that case one should have no settled view about p.

But that misstates the situation. Ankita doesn't just have two pieces of evidence; she also has the evidence that led to her initial judgment that p. We only get to describe the case in ways that make it seem symmetric if we somehow have a reason to set that initial public evidence aside. This point is well made by Kelly (2010). And the only way I can see to justify that set aside is by adopting some principle like JSE. And JSE, as we saw, is false. Moreover, JSE is equivalent, given plausible assumptions, to an internalist principle about higher-order evidence.

So however we understand peerhood, either in terms of rationality or in terms of accuracy, the arguments for conciliationism will be tied up with arguments about higher-order evidence and hence with normative externalism.

12.3. Evidence, Public and Private

In many discussions of peer disagreement, cases are presented where it is clear that the disputants have the same public evidence. It does not follow that in those cases the disputants have the same evidence tout court. Consider this simple case.

> Stars I
>
> Ankita and Bojan are wondering how many stars there are. They both have the concept of a prime number, but they aren't familiar with Euclid's proof of the infinity of primes. In fact, they both suspect, given the decreasing frequency of primes, that they run out eventually. In the course of their research into the stars, they run

> into the Delphic Oracle, who is known to always speak the truth. The Oracle says "There are as many stars as primes". Bojan takes this to be evidence that *There are infinitely many stars* is probably false. But while reflecting on it, Ankita comes up with a version of Euclid's proof that there are infinitely many primes, and concludes that there are an infinity of stars.

This is a case where Ankita should not conciliate in light of her disagreement with Bojan. She has a proof that there are infinitely many primes and Bojan does not. So she should not change her views. But that's not really a case that the most plausible form of conciliationism gets wrong. For reasons that should be familiar from previous chapters, we should treat Ankita as having more evidence than Bojan. Her reconstruction of Euclid's proof that there are infinitely many primes is a bit of evidence she has that Bojan does not.

This all suggests a very weak, and hence easier to defend, version of conciliationism. It only applies to cases where two parties have differing views about a proposition, and the following four conditions are met.

1. The two parties have no reason external to this disagreement to think that one is more likely to be rational than the other.
2. The two parties have no reason external to this disagreement to think that one is more likely to be accurate than the other.
3. The two parties have the same public evidence.
4. The two parties have the same private evidence.

The evidence aggregation theory of disagreement is anti-conciliationist in that in this extreme case, it denies that both parties should conciliate. If one party is acting rationally and the other is not, the first party should stick to their view.

But even though I'm not a conciliationist in theory, this kind of case brings out why I'm sympathetic to conciliationism in practice. These four conditions are met in vanishingly rare circumstances. And when they are not met, there are quite mundane reasons for thinking that each party should typically conciliate. A running theme through this chapter will be that the cases thought to motivate conciliationism do not satisfy these four criteria, and hence it is possible for an anti-conciliationist to consistently say that each party should move towards the others view in ordinary cases

Two more points of clarification before we move on.

First, I'm going to start by looking at a very specific form of conciliationism, namely Adam Elga's Equal Weight View (EWV). The EWV says that when two people are peers, and they have the same evidence, and they learn that they have credences c_1 and c_2 in a disputed proposition p, they should each adopt a credence half-way between their initial credences. That is, their new credence in p should be $(c_1 + c_2)/2$. The EWV is not by any means the only version of conciliationism. Indeed, it faces some difficult technical problems, described by Jehle and Fitelson (2009) and by Levinstein (2013). But as long as we are careful, we can see which objections are only problems for the EWV, and which form more general problems for conciliationism.

Second, it is very important here, as almost everywhere in epistemology, to respect the distinction Gilbert G. Harman (1986) draws between inference and implication. We can see this by looking at another example about stars.

> Stars II
>
> In this world Ankita and Bojan are very knowledgable about primes. Indeed, they are among the co-authors of that world's counterpart paper to Polymath (2014). This time the oracle tells them that there are as many stars as twin primes. Ankita infers that there are probably infinitely many stars, but it is too soon to be completely confident. Bojan, on the other hand, becomes completely certain that there are infinitely many stars.

In my opinion, and for that matter Ankita's, the evidence Ankita and Bojan have conclusively settles the question of whether there are infinitely many stars. What they know about primes, plus what they know about the oracle, plus what they are told by the oracle, probably entails that there are infinitely many stars. So there is, probably, a conclusive implication from their evidence to that conclusion. But there is no reasonable inference from their evidence to the conclusion that there are infinitely many stars. That inference requires knowing something that is not in evidence, namely that there are infinitely many twin primes. The fact that this fact is a logical truth (or at least is logically entailed by things they know about primes) is irrelevant. A probably conclusive implication can be a definitely unreasonable inference, and is in this case. Unless Bojan has a proof of the twin prime conjecture up his sleeve, one that he hasn't shared with his co-authors, he should move his credences in the direction of Ankita's. That is, he should conciliate. It's possible that Ankita should conciliate too; I haven't said nearly enough about the case to settle that one way or the other. I think the mistaken idea that entailments generate maximally strong inferences has led to

some confusion about what to say about certain cases, and that will become relevant as we progress.

12.4. Independence and Conciliationism

In early writings on concliationism, such as those by Elga (2007) and Christensen (2009), there was a line of argument from principles like Independence (as we've discussed in previous chapters) to conciliationism. This line is flawed, for reasons well set out by Errol Lord (2014). The point of this section is simply to rehearse Lord's arguments before moving onto other possible motivations for conciliationism.

There are weaker and stronger versions of the kind of Independence principle that Elga, Christense and others use. The strongest such principle says that in any dispute, a party to the dispute can only reasonably conclude that the other party is wrong based on reasons independent of their reasons for having a disputed view. But that leads to very odd predictions in cases like this.

> Bus Stop
>
> While waiting at the bus stop, Ankita is approached by Bojan , who tells her that he is certain she lives in a shoe. Ankita is fairly confident, based on long familiarity with her apartment, that she lives in an apartment, not a shoe.

Ankita doesn't have to find independent evidence that Bojan is mistaken to hold onto her belief that she lives in an apartment. Perhaps in some realistic versions of Bus Stop, Bojan would appear drunk or be slurring his words, and that would be the relevant independent evidence. But those external clues are not necessary. Bojan could appear perfectly sane and sensible in every respect except his firm belief that Ankita lives in a shoe, and she could still dismiss his view. So this strongest independence principle is false.

More plausible independence principles restrict the circumstances in which one must rely on independent reasons to dismiss a conflicting view. There are two interesting restrictions we could look at:

- Independence might be restricted to cases where the disagreeing parties are known to be just as good at reading the evidence. (We could break

this down into two sub-cases depending on whether 'good' is understood in terms of accuracy or rationality, but this won't matter.)

- Independence might be restricted to cases where the disagreeing parties are known to have the same evidence.

But, and this is the crucial point that Lord makes, neither of these restrictions on their own gives us a plausible principle. If we only impose the first restriction, we end up with the implausible conclusion that Ankita is expected to conciliate in this case.

> Party
>
> Ankita and Bojan are just as good, in both senses, at working out where a party is given some evidence. But Bojan hasn't looked at the invitation to tonight's party in weeks, so is uncertain whether the party is on State St or Main St. Ankita looked at the invitation two minutes ago, and is certain the party is on State St.

It would be absurd to think that because Ankita's credence that the party is on State St is 1, and Bojan's is 0.5, and they are just as good at working out where parties are given some evidence, that Ankita's credence that the party is on State St should move to 0.75. Rather, she should conclude that Bojan hasn't looked at the invitation recently. And she should conclude that simply because Bojan has a different credence to her about where the party is. That's what an independence principle that only imposes the first constraint would rule out, so such an independence principle is false.

Nor will the second restriction on its own do. If we restrict the restriction to public evidence, then Stars I is already a counterexample to it. But we can come up with cases where arguably Ankita and Bojan even have the same private evidence, and the restriction is still not sufficient.

> Diagnosis
>
> Ankita is a professor at a medical school, and Bojan a student. The students at her school are very good; often they are as good at diagnosis as the professors. And Bojan has done, Ankita knows, very well on his theory exams. But some students who know a lot of theory are very poor at making a diagnosis based on material in a patient's file. So Ankita pulls out a file at random for her and Bojan to look at. Given the symptoms displayed, Ankita is very confident in a particular diagnosis. But Bojan has no idea what to say about

> the case; his best guess is that we should have low but positive credence in several distinct diagnoses.

In this case, Ankita shouldn't infer that she had been over-confident. She should conclude that, despite his solid background, Bojan isn't very good at making a diagnosis. I've obviously simplified a lot, but this seems like a very natural way for professors to test whether their students have or lack a practical skill. Now perhaps the best explanation of this case is that Bojan really lacks some evidence, despite his doing well on tests. That's actually what I suspect is going on. But I suspect most people, and certainly most conciliationists, don't think that. It is much easier to motivate conciliationism if we think that there is a skill of processing evidence that goes well beyond the possession of evidence, and that in cases like this one what's happened is that Bojan lacks that skill. (Why say conciliationism is easier to motivate if one posits large skill differences that go beyond evidence possession? Because now we can say why one person should defer to another without thinking the other person has evidence they lack; the other person may have more skills.)

Now if independence just requires that the parties had the same evidence, and this is a case where the parties have the same evidence, it would be an independence violation for Ankita to infer from Bojan's lack of certainty in any diagnosis to his lack of skill in making diagnoses. Rather, she should conciliate with him, and lose confidence in her diagnosis. That's wrong, so this independence principle is too strong.

So just putting each of these restrictions on independence singularly does not yield a viable principle. What happens if we put both restrictions on at once, and say independence holds only if the parties are known to have the same evidence and known to be just as good (in some sense) at processing it? Lord points out that then we don't have a premise in an interesting argument for conciliationism. Rather, the independence principle that is supposed to motivate conciliationism has just become a statement of conciliationism. So it can't provide any independent support for it.

12.5. Circularity and Conciliationism

Conciliationism has been supported, or at least anti-conciliationist positions opposed, with arguments that anti-conciliationism lapses into an implausible kind

of circularity. Here are a couple of quotes setting out this kind of worry. First, from Adam Elga.

> To see the correctness of the equal weight view, start with a case of perceptual disagreement. You and a friend are to judge the same contest, a race between Horse A and Horse B. Initially, you think that your friend is as good as you at judging such races. In other words, you think that in case of disagreement about the race, the two of you are equally likely to be mistaken. The race is run, and the two of you form independent judgments. As it happens, you become confident that Horse A won, and your friend becomes equally confident that Horse B won.
>
> When you learn of your friend's opposing judgment, you should think that the two of you are equally likely to be correct. For suppose not–suppose it were reasonable for you to be, say, 70% confident that you are correct. Then you would have gotten some evidence that you are a better judge than your friend, since you would have gotten some evidence that you judged this race correctly, while she misjudged it. But that is absurd. It is absurd that in this situation you get any evidence that you are a better judge ...
>
> Furthermore, the above judgment of absurdity is independent of who *in fact* has done a better job. Even if in fact you have judged the series of races much more accurately than your friend, simply comparing judgments with your friend gives you no evidence that you have done so. (Elga 2007, 486–87, emphasis in original)

And second, from Diego E. Machuca. (This quote comes just after a presentation of Thomas Kelly defending something close enough, for current purposes, to the evidence aggregation view I'm defending.)

> Kelly maintains that one can be justified in thinking that one has appropriately responded to the first-order evidence even in the absence of independent evidence that one has done so. For the reason why one takes up a given belief is precisely that one *recognizes* that it is supported by the evidence one possesses, and one would not be able to recognize this if one were unjustified in thinking that the evidence does support the belief in question. I confess that I cannot see how this move is not question-begging all the way through. Just

> as one can affirm that one's opinion is justified because one recognizes that the available evidence supports it, so too one's opponent can affirm that his opinion is justified because he recognizes that the available evidence supports it. And if one were to argue that one's opponent is clearly mistaken because one would not recognize that one's belief is supported by the evidence if one were not justified in thinking that it is, one's opponent would retort that it is he who cannot be mistaken simply because he would not recognize that the evidence supports his belief were he not justified in so thinking. (Machuca 2013, 77–78)

There are two things to say about this kind of argument. The first is that the strong form of the objection Elga makes is not really a response to the evidence aggregation view, but to the view that any agent is entitled to privilege their own view over others', simply because it is their own. But the kind of reasoning Elga worries about is only available to the one who has got things right, not to both parties. So the worry is not that everyone could have their self-confidence rise, but that those who get things right could become more confident in their ability to get things right in virtue of their recent track record of having got things right. And that doesn't look like much of a worry. (I'm here not far away from the replies that Andrew Rotondo (2013) makes to circularity arguments for conciliationism.)

But at this point we run into Machuca's complaint. If the successful can be more confident in their own ability in virtue of their successes, won't those who merely think they are successful become more confident in their own ability in virtue of their own perceived success? And, in this game, doesn't everyone perceive of themselves as being successful?

There are a number of ways we could try to turn these rhetorical questions into arguments. For the reasons I went over in chapter 9, none of the resulting arguments will work. The underlying argument could be that appropriate epistemic methods must be bidirectionally luminous; everyone must be able to know if they are applying them correctly. But that kind of argument falls to the Williamsonian anti-luminosity arguments. Or it could be that appropriate epistemic methods must be sensitive, in Nozick's sense. But that kind of argument falls to the anti-sensitivity arguments. And so on for all the other ways of precifisying the argument that evidentialism licences noxiously question-begging practices.

Now I will note one sense in which those replies in chapter 9 might miss the

mark. Machuca is defending a form of Pyrrhonian scepticism. And many of my defences of externalism involved showing that the principles deployed against externalism had implausible consequences. In particular, they implied Pyrrhonian scepticism. Now that won't look implausible to a Pyrrhonian like Machuca. Here I must simply note that I'm taking it as a fixed point that we do know a lot, and that Pyrrhonian scepticism is false. This obviously loses some potential converts, but I doubt it is possible to find philosophical arguments that work for one's position against all possible rivals (Lewis 1982).

12.6. Six Examples

The last two sections reply to arguments based on general theoretical principles in favor of conciliationism. The prospects for this way of defending the EWV, or indeed any conciliatory position, look dim. But these general theoretical principles have not been what have most moved philosophers towards conciliationism. Rather, they are moved by the idea that the EWV, or at least some form of conciliationism, is the best explanation of the clear facts about some simple cases. The literature here, as with the literature on higher-order evidence, suffers from that "main cause of philosophical disease–an unbalanced diet: one nourishes one's thinking with only one kind of example." (Wittgenstein 1953, sec. 593). I can't claim to offer a balanced diet, but I can offer the start of a more varied one. Here are six new morsels that will form the basis of the discussion to follow. I'm going to argue that the evidence aggregation view can explain the last two, while conciliationist can not. And I'll argue there is no case that the conciliationist can explain while the evidence aggregation theorist can not.

12.6.1. Arithmetic

Ankita and Bojan are working on some arithmetic problems. They both know that they have a similar track record at these problems; both are reliable, with very similar rates of mistakes. They are trying to work out *What is 22 times 18?*. Ankita correctly works out that it is 396; Bojan says that it is 386. What should their credences in each answer be?

12.6.2. Jellybeans

Ankita and Bojan are trying to guess how many jellybeans are in a sealed, transparent container. They both have equal access to the container, and they both know that they have similarly good track records at this kind of game. Ankita correctly guesses that there are 396; Bojan guesses that there are 386. What should their credences in each answer be? (A similar case is considered by Jack L. Treynor (1987).)

12.6.3. Detectives

Ankita and Bojan are the two best murder detectives in the world. They both know that they are the only peers they each have, and that they have very similar track records of success, with equal (and rare) failures. They are brought in to solve a mystery that no one has made any progress on. Each quickly sees that it could only be the butler or the gardener. Bojan has equal credence in each suspect, but Ankita figures out a subtle reason that it could not have been the gardener, so is sure the butler did it. And in fact the butler did do it, and Ankita is right about why the gardener could not have done it. After they compare credences, Bojan giving equal credence to each suspect, and Ankita being sure it is the butler, what should their credences in each answer be? (I owe this case to Ben Levinstein (2013).)

12.6.4. Football

Ankita and Bojan are both very good at predicting football games of different codes. They both typically make highly rational predictions, and they both have excellent (and similar) records for accuracy. They both know all this, and they have the same public evidence about this weekend's matches. They are comparing their credences in the home team winning ahead of two big matches: an Australian Rules match in Melbourne, and an English Premier League match in London. For each match, Ankita has a credence of 0.9 that the home team will win, and Bojan has a credence of 0.1 that the home team will win. They both regard the matches as completely independent, so Ankita's credence that both home teams will win is 0.81, while Bojan's is 0.01, and each of them have credence 0.09 in each of the hypotheses that one particular home team will win and the other will not. Once they share their credences with each other, what should their credence be that (a) the home team will win the Australian Rules

match, (b) the home team will win the English Premier League match, and (c) both home teams will win?

12.6.5. Simple Arithmetic

Ankita and Bojan are working on some arithmetic problems. They both know that they have a similar track record at these problems; both are reliable, with very similar rates of mistakes. They are answering the question *What is 2 plus 2?*. Ankita says it is 4; Bojan says that it is 5. What should their credences in each answer be?

12.6.6. Doctors

Ankita and Bojan are the two best cardiologists in the world. They know each other to be peers, the only peers each has. They are brought in to diagnose a case that has stumped all the other experts in the field. Ankita judges that it is likely disease A, but she is just short of fully believing it is disease A, since she thinks disease B is an unlikely, but real, possibility. This is the rational response to the evidence. Although the patient has disease A, the evidence available to an expert cardiologist is just short of being sufficient to ground knowledge that the patient has disease A, since B is also a realistic possibility. She reports all this when she and Bojan compare notes, but Bojan reports that he is confident that the patient has disease A. What should their credence in each diagnosis be?

12.6.7. My Verdicts

These cases are, in general, not so clear that we can simply know what is true about them after a moment's thought, and use that knowledge to evaluate theories. But for the record, here are my verdicts on the cases.

In Arithmetic, I think a lot depends on the finer details of the case, particularly on how Ankita got to her answer. But I think no matter how those details are filled in, there isn't a lot of pressure on her to conciliate. Now this isn't a popular view. Much of the motivation for conciliationism comes from thinking that in versions of Arithmetic where the sum in question is not specified, there is rather strong pressure to conciliate. We'll come back to that idea several times below.

In Jellybeans, I think they clearly should conciliate. And unlike in Arithmetic, this conciliation should take the form of not just lowering their credences in their preferred answers, but in increasing their credence in answers between the two they offered. In Jellybeans, the announced answers should increase their confidence that the answer is 391, which is not what should happen in Arithmetic.

I have no idea what the answer to the third question in Football, about the appropriate credence in the compound proposition, is. We'll say a bit below about why this is such a hard question.

In each of the last three cases, Detectives, Simple Arithmetic and Doctors Ankita should not conciliate, and Bojan should move his credence dramatically in the direction of Ankita's. Or at least so I say.

12.7. Equal Weight and the Cases

On the face of it, the EWV gets at most one of the six cases right. After all, the only case where it seems even prima facie right to move to a credence half-way between the two expressed views is Arithmetic. But a more nuanced understanding of the cases lets EWV handle Jellybeans, and a more subtle version of conciliationism does well (or at least well enough) with Detectives and Football. If there is a case-based objection to conciliationism, it comes from the last two cases. But first I want to go over why the second, third and fourth cases are really not problems for conciliationism. Why, as an anti-conciliationist, should I do that? It's for two reasons. First, I want to demonstrate how hard it is to use any case around here to show that a particular view on disagreement is wrong. Second, we get an interesting insight into the range of possible and indeed plausible versions of conciliationism by working through the cases carefully.

The apparent problem with Jellybeans is that it seems the rational reaction, for both Ankita and Bojan, is to increase their credence in a particular hypothesis that neither of them endorses, namely that there are 391 beans in the jar. But it isn't hard to see that this is a merely apparent problem. What credences should we attribute to Ankita when she announces her guess of 396? Presumably not that she has credence 1 that there are 396, and credence 0 in everything else. Given what we know about jars of jellybeans, and human visual capacities, it is best to interpret her as saying that the mode of her credal distribution over the competing hypotheses about the content of the jar is 396. But that distribution will presumably be fairly spread out, and indeed fairly flat around the peak.

Similarly, Bojan will have a credal distribution that is spread out, and fairly flat around its peak of 386. If we average out those distributions, it could easily be that the peak of the new distribution is at 391. That happens, for instance, if each of Ankita and Bojan's distributions are normal distribution, with a mean at the number they announce, and a standard deviation of 10.

Now there are hard questions about how we do, or even could, know that the number they utter means that they have just this credal distribution. But that's not particularly our problem here. The question is what the parties to the dispute should do given that we add to their evidence each other's credence distribution. Questions about how we could know what another person's credence distribution are, while fascinating, are not at issue here. This is a point worth keeping in mind as we work through the examples.

The EWV does rather badly on Detectives, but other versions of conciliationism do better. Assuming that the detectives are actually very good at their jobs, then neither would have formed the conclusion that the butler did it without a very good reason. If one of them believes this, and the other does not, the one who does not should believe that they've missed a reason. So they should largely defer to the other.

Note that the reasoning in the last paragraph is entirely symmetric, and doesn't directly make use of the fact that Ankita was right to infer that it was the butler. So it is reasoning that should be available to the conciliationist, even if it isn't available to the equal weight theorist. And there is a natural method for how to get the right result in Detectives in a conciliationist-friendly way. The method in question is one I'm taking from some work by Sarah Moss (2011).

Imagine that Chika is not a detective, and has no particular expertise in solving murders. Moreover, she has very little information that bears directly on the case. What she does know is what Ankita and Bojan think; she knows that Ankita is confident the butler did it, while Bojan is uncertain. The reasoning from two paragraphs ago is available to Chika too. She can think that Ankita wouldn't be so confident unless she had a very good reason, so she can infer that it is very likely that the butler did it.

One natural form of conciliationism says that the parties to a dispute face the same normative pressures as an outsider, like Chika. Whatever is rational for Chika to do given the knowledge just of the parties' credences, and their track records and backgrounds, is rational for the parties to the dispute to do. In general, that will mean conciliating, since in general Chika should form a credence somewhere between the parties' credences. But that isn't always true. If Ankita

and Bojan were both 90% confident that it was the butler, and that's all Chika knows, then Chika should give some credence to the possibility that Ankita and Bojan have noticed independent reasons for thinking it is the butler, and should have a credence in the butler's guilt slightly higher than 0.9. Nevertheless, the view that insiders to the dispute, like Ankita and Bojan, should end up in the same place as an outsider, like Chika, who knows just the credences, seems to capture the idea at the heart of conciliationism.

I haven't said very much in general about how Chika should reason about these cases. Ben Levinstein (2013), to whom I owe this example, thinks that Chika should have a credence function that minimises the sum of Ankita and Bogan's expected inaccuracy. He persuasively argues that this method delivers the right result in a number of tricky cases.

We can also think about Football as an 'insider-outsider' problem. This case is really rather hard. I used to think it was a counterexample to any form of conciliationism, since conciliationists would have to say that each party would improperly regard the games as probabilistically dependent after learning about the disagreement. But I now think that both premises of this little argument (that conciliationism implies probabilistic dependence, and this is bad) are dubious. The case is just a hard case for everyone, and we can see that by thinking about it from Chika's perspective. (The next few paragraphs draw on work by Julia Staffel (2015).)

Assume that Chika knows nothing about football (of any code), but does know about Ankita and Bojan's predictive records and their credences concerning these games. And assume that she's an ideal aggregator. Finally assume, more or less for reductio, that Chika aggregates probabilistic judgments by taking the linear average of them. (If that's right, the EWV and the 'insider-outsider' version of conciliationism coincide; if it isn't right, they don't.) The following table gives Ankita, Bojan and Chika's credences and conditional credences, assuming that Chika does this. I'll use p for the home team wins the Australian match and q for the home team wins the English match.

	p	q	$p \wedge q$	$p \mid q$
Ankita	0.9	0.9	0.81	0.9
Bojan	0.1	0.1	0.01	0.1
Chika	0.5	0.5	0.41	0.82

The key number is in the bottom right. Assuming that Chika plans to update by

conditionalisation, that means that although her credence in q is now 0.5, if she learns p , it will rise to 0.82.

It has been argued, e.g. by Loewer and Laddaga (1985) and Jehle and Fitelson (2009) that this is a mistake for the following reason. Ankita and Bojan both take the games to be probabilistically independent. So Chika, who only has their credences to go on, should take them to be independent too. This argument doesn't work, for a reason Sarah Moss (2011) gives. The probabilities in this table are evidential probabilities. Even if the games are physically independent, it could be that the result of one gives Chika evidence about the other. And that is what happens; if she learns p she gets one more data point in favor of Ankita's general accuracy in football-predicting, and against Bojan's. So it is plausible that, for her, learning p will raise her credence in q .

What isn't plausible, as Staffel notes, is that it could raise her credence that much. We can imagine, consistent with everything I've said so far, that Chika has a lot of evidence about Ankita and Bojan's track records. If p is true, then Ankita did better at forecasting p than Bojan did. So that's a reason to no longer given exactly equal weights to their forecasts. But for all I've said so far, this might mean that we have a data set consisting of 1001 times that Ankita's forecast was better, and 1000 times that Bojan's forecast was better. That does not look like a good reason to have a probability for q that is several times closer to Ankita's forecast than it is to Bojan's. More generally, just what number goes into the bottom right of the table should be sensitive to how much information we have about Ankita and Bojan, and not just to the balance between them. Arguably, having a conditional credence for q given p of 0.82 could be reasonable if Chika knew almost nothing about Ankita and Bojan before the games were played. But it is not reasonable if she has a very substantial set of results where they have both done very well, and within that a substantial and balanced set of results in games where they have disagreed. But the Equal Weight View is insensitive to the quantity of information that Chika has.

So the Equal Weight View is wrong about this case. It gives an implausible prediction, and it is insensitive to a factor that we know to be relevant. But the failure of Equal Weight does not mean that conciliationism fails. Saying just what values should go in the two right-most boxes in the bottom row is a very very hard question. But presumably it has at least one good answer. It doesn't seem like this is an epistemic dilemma for Chika. So the conciliationist can still say something substantive about how Ankita and Bojan should react to learning about each other's forecast. The conciliationist, I suggest, should say that Ankita and Bojan should adopt whatever credences Chika should adopt. This

is a substantive and interesting claim. I think it is false, but I don't think it is obviously false. Ideally the conciliationist who says this would have something a little more substantive to say about what Chika's credence should be. But ideally any epistemologist who discusses the problem would have something a little more substantive to say about what Chika's credence should be, so this isn't a particular problem for conciliationism. Nor is there any reason to think that adopting conciliationism makes it harder to say what Chika should do. So while this looks like a hard case, I don't think it can be used in an argument against conciliationism. Or, at least, it can't be so used just yet. Perhaps we could solve the problem of how Chika should react, and then show it is implausible for Ankita and/or Bojan to react that way. But I'm not in any position to run that argument, because I don't know what Chika should do.

What I'm saying here is very similar to what I said in the chapter 6 about the problem of inter-theoretic value comparisons. In both cases, normative internalism makes vivid a particularly hard epistemic problem. But the problem in question, in each case a problem about aggregation, was hard to start with, and isn't any harder in virtue of internalism. The fact that internalism makes the problem vivid is not in itself a reason to reject internalism.

On the other hand, Doctors is a problem for conciliationism, and looking at the problem through Chika's eyes doesn't help the conciliationist. If Chika knows that Bojan is certain of a diagnosis, and that Ankita gives that credence a very high credence just short of belief, it seems prima facie plausible that Chika should conclude from that that the diagnosis is correct. Unless we have some way to motivate a theory of judgment aggregation where the aggregate opinion is never more confident in a proposition than the weakest member, there must be some such cases where Chika should believe the diagnosis is correct. But Ankita should not share this confidence. She should not find her doubts assuaged by Bojan's not sharing them. So Doctors is a counterexample to the 'insider-outsider' version of conciliationism. And that's the only version that seems to get Football right. So no version of conciliationism can get both these cases right.

It is easy for the evidentialist to say what's going on in Simple Arithmetic. Ankita has maximally strong evidence that 2 plus 2 is in fact 4. That's not just because the conclusion is a logical truth. There are plenty of logical truths that we have insufficient evidence to believe, either because we don't know which logics validate them, or because we don't know what the correct logic is. Rather, it is because the inference from $x = 2 + 2$ to $x = 4$ is one that is immediately justified, without the need for further steps. Bojan's disagreement can't dislodge that.

But how can the conciliationist handle the case? It doesn't seem very plausible to say that when an otherwise reasonable person says that two plus two is five, we're obliged to doubt that it is four. The usual response on behalf of conciliationists is to appeal to the notion of 'personal information'. The idea was first developed by Jennifer Lackey (2010), but I want to first mention the version of this defence put forward by David Christensen (2011). (Christensen is describing a scenario where the narrator plays the role of Ankita, and Bojan is their friend.)

> If such a bizarre situation were actually to occur, I think one would reasonably take it as extremely unlikely that one's friend (a) was feeling as clear-headed as oneself; (b) had no memories of recent drug-ingestions or psychotic episodes; and most importantly, (c) was being completely sincere. Thus, to use Lackey's term, one's personal information (that one was feeling clear, lacked memories suggesting mental malfunction, and was being sincere in one's assertion) would introduce a relevant asymmetry, and one could reasonably maintain one's belief.

The first thing to be said here is that (c), which is what Christensen adds to Lackey's original characterisation, is beside the point. The question is what Ankita should do given that Bojan believes that 2 plus 2 is 5. It's not the separate question of whether she should believe he believes that, given his utterance. So questions of sincerity are beside the point. Then the question is whether (a) and (b), which are the aspects of personal information that Lackey originally highlighted, are enough to help.

And it is hard to see how they could be. If the reason for discounting Bojan's opinion rested on one's personal information, then the more information we get about Bojan, the more worried we should be. But I rather doubt that running a drug test on Bojan, to see whether (b) is a relevant difference between him and Ankita, should make any difference at all to Ankita's confidence.

More generally, this explanation rests on an odd view about epistemic capacities. Ankita's ability to do simple arithmetic is not, according to Christensen, a sufficient ground to believe that two plus two is four. But her ability to detect differences in capacities and aptitudes between two people, one of whom is herself, is enough of a ground. Speaking personally, I'm sure I'm much better at simple arithmetic than I am at doing such comparisons. Indeed, my abilities to make such comparisons intuitively are so weak that I could only possibly do them by careful statistical analysis, and that would require, among other things, being able to add two plus two. In other words, if I can't know what two and two is, I

can't process the evidence that might tell for or against the abilities of one party or another. So the conciliationist doesn't have a good explanation of how we can hold on to knowledge in these simple cases.

Simple arithmetic cases are important not just because they raise problems for conciliationism, but because they tell us something about what's at issue in debates about disagreement. Consider this argument by David Enoch for thinking that in debates about disagreement as such, we should treat the parties to the disagreement symmetrically.

> Second, our question, as you will recall, was the focused one about the epistemic significance of the disagreement itself. The question was not that of the overall epistemic evaluation of the beliefs of the disagreeing peers. Kelly is right, of course, that in terms of overall epistemic evaluation (and barring epistemic permissiveness) no symmetry holds. But from this it does not follow that the significance of the disagreement itself is likewise asymmetrical. Indeed, it is here that the symmetry is so compelling. The disagreement itself, after all, plays a role similar to that of an omniscient referee who tells two thinkers 'one of you is mistaken with regard to p'. It is very hard to believe that the epistemically responsible way to respond to such a referee differs between the two parties. And so it is very hard to believe that the epistemic significance of the disagreement itself is asymmetrical in anything like the way Kelly suggests. (Enoch 2010, 657)

Well, consider the case when p is the proposition that two plus two is four, and Ankita is the party who believes p, while Bojan rejects it. Having an omniscient referee tell the parties that one of them is mistaken should produce asymmetric responses in the two parties. Now maybe there are only a small class of cases where this is the case, and what Enoch says is right in the majority of cases. But we can't argue for that on perfectly general grounds about the nature of disagreement, because it fails in extreme cases like Simple Arithmetic. The argument that it holds in normal cases needs a distinct defence.

12.8. The Evidence Aggregation Approach

Having gone over how conciliationism handles, or doesn't handle, the cases, let's compare it to how an evidence aggregation view handles them. We'll look at

them in reverse order, because the earlier cases are harder for the view.

Evidence Aggregation gets Simple Arithmetic right. Ankita has clear and compelling evidence that two plus two is four. The fact that two plus two is four is part of her evidence, and when the conclusion is part of one's evidence, that is maximally strong support. Learning that something has gone badly wrong with Bojan's arithmetic competence or performance does not make her lose this evidence.

It also gets Doctors right, by treating the case as parallel to the case of Roshni from chapter 8. When Ankita learns that Bojan is very confident that the patient has disease A, that isn't yet evidence that Bojan has stronger evidence that the patient has disease A. It might mean simply that Bojan hasn't considered the possibility of B, or that he has overly hastily dismissed it. And indeed, that's just what has happened. Until Ankita learns why Bojan has the credences he does, she can reasonably, if provisionally, keep her current credences. After all, there may not be any new evidence in favour of diagnosing A. And when she does learn why Bojan has these credences, she should stick to her initial view. That's not because it was her view, but rather because it was the view best supported by the current evidence.

From this perspective, Detectives is just like Doctors. When Ankita hears Bojan's credence, it is reasonable for her to infer that she has some evidence that Bojan lacks. This evidence need not be public evidence; it might be more like the kind of evidence a mathematician gets when working through a proof. But it is reasonable for her to infer, given just the facts about their conflicting credences, that Bojan has simply missed the reason that it must have been the butler. So she doesn't have new evidence that it wasn't the butler, so her credence shouldn't move.

The last two cases are not like most everyday cases of disagreement. The usual situation, when another person disagrees with us, is that they have evidence we lack. Or, at least, it is usually the case that one should give the possibility that the other person has extra evidence substantial credence. That's why it is usually the case that one should conciliate. The default view is that the other probably has good evidence we lack, and that is reason to move one's attitude towards the other's. It is very hard to say in general when one should abandon this default stance. Indeed, it is very hard to even say whether the 'should' in question is moral or epistemic. It feels like an epistemic question at first, but perhaps moral considerations to do with humility, respect and friendship are also relevant factors. But we shouldn't let the fact that it is hard to give a general

theory here prevent us from saying something about some cases. And we should say that Doctors and Detectives are among the (presumably rare) cases where one party, in this case Ankita, is warranted in holding firm to their beliefs.

It's a little harder to know what the evidence aggregation view should say about Football. The case as presented didn't include much detail about how Ankita or Bojan came to their conclusions. If I was in one or other of their positions, I would likely infer that the other had picked up on some reason I missed, but also that they had probably missed some reason I'd seen. So I would be tempted to conciliate, because this is a case where the conflicting credences really are useful evidence that there is (private) evidence that would motivate a change of view.

While the evidence aggregation view doesn't have a firm theoretical recommendation, it does have a firm practical recommendation. Each party should ask the other why they have the view that they do. Assuming it is possible to ask the other this question, and the disagreement is about something significant enough to make it worth the bother, this is pretty much always the practical recommendation. As far as I can tell, intuition and folk wisdom agree with the evidence aggregation view on this point. And it is hard to see how rival views of disagreement could motivate such a strong recommendation to ask the other person "Why do you think that?". After all, those rival views already say what the disagreeing parties should do, and the answer is not sensitive to why the other person has the views they do. If it turns out that all the reasons Bojan can offer are ones that Ankita had already properly weighed, she should revert to her initial credence. But probably he has thought of something she missed, and probably she has thought of something he missed, and adding those reasons together will bring their views closer together.

The conciliationist thinks that Ankita and Bojan should aggregate the outputs of their deliberation. The evidence aggregation view says that they should aggregate the inputs to their deliberation. If the only evidence they have as to those inputs is the outputs, then they should use the outputs to make reasonable guesses as to the nature of the inputs, and aggregate them. But this is very much a second-best solution; the best thing to do is to find out exactly what the inputs were. That is exactly what good interlocutors do. The primary reaction to hearing that someone has a very different view to one's own shouldn't be to jump to a new credence, it should be to find out why they have the conflicting view.

In Football it was plausible that the two parties would have different evidence; in Jellybeans it is just about certain. Ankita and Bojan will have had different

appearances when they looked at the jar, they will have seen it from different angles, they will be bringing different histories with these kinds of estimation tasks to bear on the subject, and so on. In Football it was likely that the parties will have different views about the question because they have different evidence; in Jellybeans it is practically certain. So the evidence aggregation view says, along with intuition, that this is a case where they should conciliate. It has a simpler explanation as to why their credence in hypotheses like 391 should increase than the conciliationist offers, but both parties get to the right result for plausible reasons.

The case that's left is Arithmetic. This case seems to be the one that moves people to reject evidentialist views. Cases like Arithmetic are used as a primary motivation for conciliationist views of disagreement in, for example (Bogardus 2009; Matheson 2009; Carey 2011; Kraft 2012; Lee 2013; Vavova 2014; Worsnip 2014; Mogensen 2016) and Ebeling (2017). In many of these papers, intuitions about cases like Arithmetic are the sole motivation offered for conciliationism, or are offered as a sufficient reason to believe conciliationism. Worship says that cases like Arithmetic show that views like evidence aggregation are "not even slightly plausible" (Worsnip 2014, 6). Although cases like Arithmetic are commonly used by conciliationist philosophers, none of them ever say just what arithmetic problem is under dispute in their version of the case. The usual methodology s to describe the kind of arithmetic problem at issue, then present the conflicting answers that the peers give. I'm using a more concrete example because my analysis turns on being able to talk about the particular arithmetic problem under discussion.

The first thing to note about Arithmetic as I've presented it is that it leaves out some details about how Ankita came to her conclusion. (Remember that the versions offered in the literature are even lighter on details.) So I'll go over two variants of the case. The variations will be important enough that I'll introduce new characters to participate in them. Each character has the same prior relationship to Bojan as Ankita does.

Deanna thinks to herself that 22 times 18 is 20 times 18 plus 2 times 18, so it is 360 plus 36, so it is 396. That strikes her as conclusive, so she announces that it is 396. Bojan then says he thinks 22 times 18 is 386. So Deanna decides to double check. She thinks that 22 times 18 is 20 plus 2 times 20 minus 2, so it is 20 squared minus 2 squared, so it is 400 minus 4, so it is 396. She now feels confident sticking to her original verdict.

Efrosyni thinks to herself that 22 times 18 is 20 times 18 plus 2 times 18, so it is

360 plus 36, so it is 396. But she feels she should double check. So she thinks that 22 times 18 is 20 plus 2 times 20 minus 2, so it is 20 squared minus 2 squared, so it is 400 minus 4, so it is 396. She now feels confident sticking to her original verdict. She then hears Bojan say that he thinks 22 times 18 is 386.

Whatever one's view about how confident Deanna and Efrosyni should end up being in their verdict that 22 times 18 is 396, they should be equally confident. After all, they have exactly the same evidence for and against it: two calculations that point to 396, and Bojan's announcement of 386. But no form of conciliationism can deliver that result. After all, conciliationism requires that a form of independence hold.[4] The reasoning that led to one's disagreeing views cannot be used to 're-check' that those views are correct. So once Efrosyni hears Bojan's disagreement, she can't rely on either of the two routes to the conclusion that she used. But Deanna is free to use the second calculation she did as independent evidence that Bojan is wrong. So the standard conciliationist has to say, falsely, that Deanna and Efrosyni should have different credences in the proposition that 22 times 18 is 396, or, equally falsely, that Deanna doesn't get any extra reason to believe that 22 times 18 is 396 when she does the double-check.

The evidence aggregation theory suggests a better analysis of the case. Consider the state of mind that Efrosyni was in when she thought, "I'd better double check this." She actually had conclusive, entailing, evidence that 22 times 18 was 396. Of course, everyone has just the same evidence at all times, so perhaps that isn't so important. What is more important is that after doing the first calculation she had evidence that a reasonable person could, other things being equal, base a belief on. Deanna was not unreasonable when she made her announcement, but yet Efrosyni in a similar position thought she should get more information. How should we explain that? We could treat this as a case where a kind of permissivism is right; Deanna was being reasonable in ending inquiry, and Efrosyni was reasonable in not ending it, despite their being in identical positions. But it is better to treat these cases as not quite identical. Efrosyni had a nagging doubt, which Deanna did not have. Perhaps that is the difference; the calculations they had both done are sufficient to end inquiry in the absence of positive reasons to extend inquiry. A nagging doubt like Efrosyni had is reasonable, and if one has such a doubt, one has a reason to address it. But it is also reasonable to not have such a doubt.

[4]The discussion of Lord's work above wasn't meant to undermine that claim. The result of that discussion was that conciliationism is equivalent to the strongest plausible independence principle, so that principle can't be used to independently defend conciliationism. That's all consistent with saying that conciliationism requires an independence principle.

If that story is right, then the evidence aggregation theorist can easily say what's going on in Arithmetic. If Ankita is like Deanna, then the exchange with Bojan provides a good reason to recheck her calculations. The idea here is that the evidence Ankita acquired by doing the calculation is good enough to close inquiry, but only in the absence of positive reason to keep the inquiry going. That reason could be internal, a nagging doubt, or it could be external, such as peer disagreement. So it is fine for an evidential aggregation theorist to say that Deanna (or Ankita if she is like her) should not necessarily conciliate, but should re-open inquiry. If Ankita is like Efrosyni, then the evidence aggregation theorist can't make that move. But she shouldn't want to. After all, Efrosyni has just as good reason to believe 22 times 18 is 396 as Deanna did after rechecking. So she has excellent reason to believe that 22 times 18 is 396. So she should keep believing it. It is much more plausible that Bojan made a rare mistake than that she made distinct mistakes on distinct calculations that ended up at the same point.

As stated, Arithmetic is not detailed enough for us to know what Ankita should do or believe. The advantage of the evidence aggregation theory is that it can explain why the missing details matter. Probably the most intuitive way to fill in the details in the original case is to make Ankita like Deanna; she does the calculation once, and easily could have a reason to double-check, but does not do this. In this case we can say Bojan's disagreement should prompt Ankita to double-check. So we can explain the case that was meant to be the best case for conciliationism. And, if one thinks the differences between Deanna's case and Efrosnyi's case needs to be explained, the evidence aggregation theory can explain them even more smoothly than the conciliationist can. So there is no argument from intuitions about cases for conciliationism, and if any side is favoured by considerations about cases, it is the evidence aggregation theory.

13. Epilogue

I've argued at length against the idea that conformity to one's own principles is a core part of ethics or epistemology. One should conform to good principles. If one's own principles are good, then one should conform to them. But that's because they are good, not because they are one's own.

One running theme of the book has been that the idea that we should conform to our principles leads to regresses. Philosophers like the idea that people should conform to their own principles because this often provides more useful, more actionable, advice than the idea that people should do what is right. But it isn't always more useful. Just as one might not know what the true principles are, one might not know how to apply principles one has chosen or adopted. So even if what matters is conformity to one's own principles, we can have disputes over who lives up to that standard. And if we want people to only be bound by constraints they can appreciate, indeed if that's why we thought conformity to one's own principles was so important, then we'll have to say that what matters is conformity to one's own judgment about what one's own principles requires. And now we're past the point at which subjectivism becomes implausible.

Another theme has been that the internalist wants beliefs to play philosophical roles that only desires are fit to play. The prudent person will perform acts that they believe will have consequences that they actually desire. They won't, in general, perform acts that they believe will have consequences that they believe they desire, or that they believe to be desirable, or that they believe to be valuable. All of these theories of prudence have only beliefs, and not desires, determine what is a prudent act, and hence are vulnerable to the technical objections to desire as belief theories. The moral person will desire things that are actually good. I think this means they will have a vast plurality of desires: to treat others with respect, to promote the general good, to keep their promises and contracts, and so on. What makes them moral is not that they have one desire, to do the good, plus some beliefs about what the good consists in; it is that they have the right desires. Similarly, the rational person will not have just one inferential disposition: to move from *it is rational to* ϕ *in my circumstances* to realizing

ϕ. There is an internalist picture that this, plus very rich beliefs about what rationality consists in, is all the rational person needs. But this is not how the rational person should operate. Rather, they will have any number of distinct dispositions, corresponding to the various ways in which rationality requires one to react to different situations.

A final theme is more ironic. Internalism is often promoted as the theory that gives us moderation and caution. Some internalists in ethics describe their view as 'moral hedging'. Internalism in epistemology is motivated by cases like Christensen's medical resident, and disagreeing peers moving from extreme views to suspension of judgment. But nothing in the internalist's theory entails that they will always be on the side of moderation and caution. Indeed, a running theme of this part of the book has been the epistemological internalist will end up taking the extreme position in any number of cases. And internalism in ethics only makes sense if you think the good agent has as a primary aim, or perhaps as a sole aim, to do what is right by their own lights. And that is not a recipe for moderation and caution. Rather, it is the characteristic of that most immoderate and reckless figure: the self-righteous ideologue.

I don't object to aiming for caution and moderation in one's theory. But a lesson of the examples we've thought about in this book is that this must be inserted in first-order theory, not as the internalist wants to do in the meta-theory. My first-order suggestions are that we are thoroughly pluralist in our theory of value, and allow that mathematical investigation is a way of acquiring evidence, not processing it. But I'm less committed to those particular suggestions than I am to the view that imitating an ideologue is a bad way to promote moderation.

Amia Srinivasan ends her excellent paper "Normativity without Cartesian Privilege" by noting that her view, one that I'd call externalitist, "invites us to return to a more tragic outlook of the normative." (Srinivasan 2015a, 287). But that tragic outlook, she argues, can be beneficial; it helps focus on injustices in practice rather than injustices in theory.

The worldview motivating this book is very similar. Reflection on what makes tragic figures tragic is a good way to appreciate this worldview. (There is a reason I started this book by quoting Shakespeare.) And the misguided ideologue, the person who governs their thoughts and deeds by the theory they think is right, but in fact is off in one key respect, is one of the great tragic figures of modernity. What might have been a minor flaw in an average person becomes, in the ideologue, a character defining vice.

We should avoid that tragic end. We should try to live well and, if our minds turn to theory, we should try to have true beliefs about what it is to live well. If all goes perfectly, there will be a pleasing harmony between how we live and how we think one should live. But aiming for that harmony is dangerous, and changing our lives to guarantee it can bring more harm than good. And we should reject philosophical theories that draw conclusions about morality or rationality from giving that harmony too exalted a place.

References

Adams, Robert Merrihew. 1985. "Involuntary Sins." *Philosophical Review* 94 (1): 3–31. https://doi.org/10.2307/2184713.

Adler, Jonathan E. 2002. "Akratic Believing?" *Philosophical Studies* 110: 1–27. https://doi.org/10.1023/A:1019823330245.

Alexander, David. 2011. "In Defence of Epistemic Circularity." *Acta Analytica* 26: 223–41. https://doi.org/10.1007/s12136-010-0100-2.

Allais, M. 1953. "Le Comportement de l'homme Rationnel Devant Le Risque: Critique Des Postulats Et Axiomes de l'ecole Americaine." *Econometrica* 21 (4): 503–46. https://doi.org/10.2307/1907921.

Arntzenius, Frank. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy* 100 (7): 356–70. https://doi.org/10.5840/jphil2003100729.

Arpaly, Nomy. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, Nomy, and Timothy Schroeder. 2014. *In Praise of Desire*. Oxford: Oxford University Press.

Ballantyne, Nathan, and E. J. Coffman. 2011. "Uniqueness, Evidence and Rationality." *Philosophers' Imprint* 11: 1–13. http://hdl.handle.net/2027/spo.3521354.0011.018.

———. 2012. "Conciliationism and Uniqueness." *Australasian Journal of Philosophy* 90: 657–70. https://doi.org/10.1080/00048402.2011.627926.

Barnett, David James. 2014. "What's the Matter with Epistemic Circularity?" *Philosophical Studies* 171 (2): 177–205. https://doi.org/10.1007/s11098-013-0261-0.

———. 2015. "Is Memory Merely Testimony from One's Former Self?" *Philosophical Review* 124 (3): 353–92. https://doi.org/10.1215/00318108-2895337.

Basu, Rima, and Mark Schroeder. forthcoming. "Doxastic Wrongings." In *Pragmatic Encroachment in Epistemology*, edited by Brian Kim and Matthew McGrath. Routledge.

Bogardus, Tomas. 2009. "A Vindication of the Equal-Weight View." *Episteme* 6 (3): 324–35. https://doi.org/10.3366/E1742360009000744.

Boghossian, Paul. 2003. "Blind Reasoning." *Proceedings of the Aristotelian Society, Supplementary Volume* 77 (1): 225–48. https://doi.org/10.1111/1467-

8349.00110.

BonJour, Laurence, and Ernest Sosa. 2003. *Epistemic Justification: Internalism Vs. Externalism, Foundations Vs. Virtues*. Great Debates in Philosophy. Malden, MA: Blackwell.

Bostrom, Nick. 2003. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–55. https://doi.org/10.1111/1467-9213.00309.

Brown, Campbell. 2011. "Consequentialize This." *Ethics* 121 (4): 749–71. https://doi.org/10.1086/660696.

Buchak, Lara. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

———. 2014. "Belief, Credence and Norms." *Philosophical Studies* 169 (2): 285–311. https://doi.org/10.1007/s11098-013-0182-y.

Burns, Jeffrey M, and Russell H Swerdlow. 2003. "Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Apraxia Sign." *Archives of Neurology* 60 (3): 437–40. https://doi.org/10.1001/archneur.60.3.437.

Calhoun, Cheshire. 1989. "Responsibility and Reproach." *Ethics* 99 (2): 389–406. https://doi.org/10.1086/293071.

Cappelen, Herman, and Josh Dever. 2014. *The Inessential Indexical*. Oxford: Oxford University Press.

Carey, Brandon. 2011. "Possible Disagreements and Defeat." *Philosophical Studies* 155 (3): 371–81. https://doi.org/10.1007/s11098-010-9581-5.

Carroll, Lewis. 1895. "What the Tortoise Said to Achilles." *Mind* 4 (14): 278–80. https://doi.org/10.1093/mind/iv.14.278.

Cho, In-Koo, and David M. Kreps. 1987. "Signalling Games and Stable Equilibria." *The Quarterly Journal of Economics* 102 (2): 179–221. https://doi.org/10.2307/1885060.

Christensen, David. 2005. *Putting Logic in Its Place*. Oxford: Oxford University Press.

———. 2007a. "Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals." *Oxford Studies in Epistemology* 2: 3–31.

———. 2007b. "Epistemology of Disagreement: The Good News." *Philosophical Review* 116 (2): 187–217. https://doi.org/10.1215/00318108-2006-035.

———. 2009. "Disagreement as Evidence: The Epistemology of Controversy." *Philosophy Compass* 4 (5): 756–67. https://doi.org/10.1111/j.1747-9991.2009.00237.x.

———. 2010a. "Higher-Order Evidence." *Philosophy and Phenomenological Research* 81 (1): 185–215. https://doi.org/10.1111/j.1933-1592.2010.00366.x.

———. 2010b. "Rational Reflection." *Philosophical Perspectives* 24: 121–40. https://doi.org/10.1111/j.1520-8583.2010.00187.x.

———. 2011. "Disagreement, Question-Begging and Epistemic Self-Criticism."

Philosophers' Imprint 11 (6): 1–22. http://hdl.handle.net/2027/spo.3521354.0011.006.

———. 2016. "Conciliation, Uniqueness and Rational Toxicity." *Noûs* 50 (3): 584–603. https://doi.org/10.1111/nous.12077.

Coady, C. A. J. 1995. *Testimony: A Philosophical Study*. Oxford: Clarendon Press.

Coates, Allen. 2012. "Rational Epistemic Akrasia." *American Philosophical Quarterly* 49 (2): 113–24.

Cohen, Stewart. 1986. "Knowledge and Context." *The Journal of Philosophy* 83: 574–83. https://doi.org/10.2307/2026434.

———. 2002. "Basic Knowledge and the Problem of Easy Knowledge." *Philosophy and Phenomenological Research* 65 (2): 309–29. https://doi.org/10.1111/j.1933-1592.2002.tb00204.x.

———. 2005. "Why Basic Knowledge Is Easy Knowledge." *Philosophy and Phenomenological Research* 70 (2): 417–30. https://doi.org/10.1111/j.1933-1592.2005.tb00536.x.

———. 2013. "A Defence of the (Almost) Equal Weight View." In *The Epistemology of Disagreement: New Essays*, edited by David Christensen and Jennifer Lackey, 98–117. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199698370.003.0006.

Conee, Earl. 1992. "The Truth Connection." *Philosophy and Phenomenological Research* 52 (3): 657–69. https://doi.org/10.2307/2108213.

DeRose, Keith. 1995. "Solving the Skeptical Problem." *Philosophical Review* 104: 1–52. https://doi.org/10.2307/2186011.

Douven, Igor. 2009. "Uniqueness Revisited." *American Philosophical Quarterly* 46: 347–61.

Dretske, Fred. 2005. "Is Knowledge Closed Under Known Entailment? The Case Against Closure." In *Contemporary Debates in Epistemology*, edited by Matthias Steup and Ernest Sosa, 13–26. Malden, MA: Blackwell.

Ebeling, Martin. 2017. *Conciliatory Democracy: From Deliberation Toward a New Politics of Disagreement*. New York: Palgrave Macmillan.

Egan, Andy, and Adam Elga. 2005. "I Can't Believe I'm Stupid." *Philosophical Perspectives* 19 (1): 77–93. https://doi.org/10.1111/j.1520-8583.2005.00054.x.

Elga, Adam. 2007. "Reflection and Disagreement." *Noûs* 41 (3): 478–502. https://doi.org/10.1111/j.1468-0068.2007.00656.x.

———. 2008. "Lucky to Be Rational."

———. 2010. "How to Disagree about How to Disagree." In *Disagreement*, edited by Ted Warfield and Richard Feldman, 175–87. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.003.0008.

Elizabeth, Princess of Bohemia, and René Descartes. 2007. *The Correspondence Between Princess Elizabeth of Bohemia and René Descartes*. Translated by Lisa Shapiro. Chicago: University of Chicago Press.

Enoch, David. 2010. "Not Just a Truthometer: Taking Oneself Seriously (but Not Too Seriosuly) in Cases of Peer Disagreement." *Mind* 119 (476): 953–97. https://doi.org/10.1093/mind/fzq070.

Fantl, Jeremy, and Matthew McGrath. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Feldman, Richard. 2007. "Reasonable Religious Disagreements." In *Philosophers Without Gods: Meditations on Atheism and the Secular*, 194–214. Oxford: Oxford University Press.

Field, Claire. forthcoming. "It's OK to Make Mistakes: Against the Fixed Point Thesis." *Episteme*, forthcoming. https://doi.org/10.1017/epi.2017.33.

Finnis, John. 2011. *Natural Law and Natural Rights*. Second. Oxford: Oxford University Press.

FitzPatrick, William J. 2008. "Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge." *Ethics* 118 (4): 589–613. https://doi.org/10.1086/589532.

Fricker, Miranda. 2010. "The Relativism of Blame and Williams's Relativism of Distance." *Aristotelian Society Supplementary Volume* 84 (1): 151–77. https://doi.org/10.1111/j.1467-8349.2010.00190.x.

Fumerton, Richard. 2010. "You Can't Trust a Philosopher." In *Disagreement*, edited by Ted Warfield and Richard Feldman, 91–110. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.003.0006.

Galbraith, John Kenneth. 1964. "Let Us Begin: An Invitation to Action on Poverty." *Harper's*, 16–26.

Ganson, Dorit. 2008. "Evidentialism and Pragmatic Constraints on Outright Belief." *Philosophical Studies* 139 (3): 441–58. https://doi.org/10.1007/s11098-007-9133-9.

Gendler, Tamar Szabó. 2000. "The Puzzle of Imaginative Resistance." *Journal of Philosophy* 97 (2): 55–81. https://doi.org/10.2307/2678446.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23 (6): 121–23. https://doi.org/10.2307/3326922.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast*. Cambridge: Harvard University Press.

Graham, Peter A. 2014. "A Sketch of a Theory of Moral Blameworthiness." *Philosophy and Phenomenological Research* 88 (2): 388–409. https://doi.org/10

.1111/j.1933-1592.2012.00608.x.

Greaves, Hilary, and Toby Ord. 2017. "Moral Uncertainty about Population Axiology." *Journal of Ethics and Social Philosophy* 12 (2): 135–67. https://doi.org/10.26556/jesp.v12i2.223.

Greco, Daniel. 2014. "A Puzzle about Epistemic Akrasia." *Philosophical Studies* 167: 201–19. https://doi.org/10.1007/s11098-012-0085-3.

Guerrero, Alexander. 2007. "Don't Know, Don't Kill: Moral Ignorance, Culpability and Caution." *Philosophical Studies* 136 (1): 59–97. https://doi.org/10.1007/s11098-007-9143-7.

Gustafsson, Johan E., and Olle Torpman. 2014. "In Defence of My Favorite Theory." *Pacific Philosophical Quarterly* 95 (2): 159–74. https://doi.org/10.1111/papq.12022.

Harman, Elizabeth. 2011. "Does Moral Ignorance Exculpate?" *Ratio* 24 (4): 443–68. https://doi.org/10.1111/j.1467-9329.2011.00511.x.

———. 2015. "The Irrelevance of Moral Uncertainty." *Oxford Studies in Metaethics* 10: 53–79. https://doi.org/10.1093/acprof:oso/9780198738695.003.0003.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: Bradford.

Hawthorne, John, and Ofra Magidor. 2009. "Assertion, Context, and Epistemic Accessibility." *Mind* 118 (470): 377–97. https://doi.org/10.1093/mind/fzp060.

———. 2011. "Assertion and Epistemic Opacity." *Mind* 119 (476): 1087–1105. https://doi.org/10.1093/mind/fzq093.

Hawthorne, John, and Amia Srinivasan. 2013. "Disagreement Without Transparency: Some Bleak Thoughts." In *The Epistemology of Disagreement: New Essays*, edited by David Christensen and Jennifer Lackey, 9–30. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199698370.003.0002.

He, Zijing, Matthias Bolz, and Renée Baillargeon. 2011. "False-Belief Understanding in 2.5-Year-Olds: Evidence from Violation-of-Expectation Change-of-Location and Unexpected-Contents Tasks." *Developmental Science* 14 (2): 292–305. https://doi.org/10.1111/j.1467-7687.2010.00980.x.

Hedden, Brian. 2016b. "Does MITE Make Right? On Decision-Making Under Normative Uncertainty." *Oxford Studies in Metaethics* 11: 102–28. https://doi.org/10.1093/acprof:oso/9780198784647.001.0001.

———. 2016a. "Does MITE Make Right? On Decision-Making Under Normative Uncertainty." *Oxford Studies in Metaethics* 11: 102–28. https://doi.org/10.1093/acprof:oso/9780198784647.001.0001.

Holton, Richard. 1999. "Intention and Weakness of Will." *The Journal of Philos-*

ophy 96 (5): 241–62. https://doi.org/10.2307/2564667.

———. 2014. "Intention as a Model for Belief." In *Rational and Social Agency: Essays on the Philosophy of Michael Bratman*, edited by Manuel Vargas and Gideon Yaffe, 12–37. Oxford: Oxford University Press.

Hookway, Christopher. 2001. "Epistemic Akrasia and Epistemic Virtue." In *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, edited by Abrol Fairweather and Linda Trinkaus Zagzebski, 178–99. Oxford: Oxford University Press.

Horowitz, Sophie. 2014. "Epistemic Akrasia." *Noûs* 48 (4): 718–44. https://doi.org/10.1111/nous.12026.

Hurley, Susan. 1989. *Natural Reasons*. Oxford: Oxford University Press.

Ichikawa, Jonathan. 2009. "Explaining Away Intuitions." *Studia Philosophica Estonica* 2 (2): 94–116. https://doi.org/10.12697/spe.2009.2.2.06.

Ichikawa, Jonathan, and Benjamin Jarvis. 2009. "Thought-Experiment Intuitions and Truth in Fiction." *Philosophical Studies* 142 (2): 221–46. https://doi.org/10.1007/s11098-007-9184-y.

Jackson, Frank. 1977. *Perception: A Reprsentative Theory*. Cambridge: Cambridge University Press.

———. 1987. *Conditionals*. Blackwell: Oxford.

———. 1991. "Decision Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101: 461–82. https://doi.org/10.1086/293312.

Jeffrey, Richard C. 1983. *The Logic of Decision*. 2nd ed. Chicago: University of Chicago Press.

Jehle, David, and Branden Fitelson. 2009. "What Is the 'Equal Weight View'?" *Episteme* 6 (3): 280–93. https://doi.org/10.3366/E1742360009000719.

Keller, Simon. 2009. "Welfare as Success." *Noûs* 43 (4): 656–83. https://doi.org/10.1111/j.1468-0068.2009.00723.x.

Kelly, Thomas. 2005. "The Epistemic Significance of Disagreement." *Oxford Studies in Epistemology* 1: 167–96.

———. 2010. "Peer Disagreement and Higher Order Evidence." In *Disagreement*, edited by Ted Warfield and Richard Feldman, 111–74. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226078.003.0007.

Klein, Peter. 2015. "Skepticism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2015. Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/archives/sum2015/entries/skepticism/.

Kolodny, Niko. 2005. "Why Be Rational?" *Mind* 114 (455): 509–63. https://doi.org/10.1093/mind/fzi509.

Kraft, James. 2012. *The Epistemology of Religious Disagreement: A Better Understanding*. New York: Palgrave Macmillan.

Lackey, Jennifer. 2010. "What Should We Do When We Disagree." *Oxford Studies in Epistemology* 3: 274–93.

Lam, Barry. 2011. "On the Rationality of Belief-Invariance in Light of Peer Disagreement." *Philosophical Review* 120: 207–45. https://doi.org/10.1215/00318108-2010-028.

Lasonen-Aarnio, Maria. 2010a. "Is There a Viable Account of Well-Founded Belief." *Erkenntnis* 72 (2): 205–31. https://doi.org/10.1007/s10670-009-9200-z.

———. 2010b. "Unreasonable Knowledge." *Philosophical Perspectives* 24: 1–21. https://doi.org/10.1111/j.1520-8583.2010.00183.x.

———. 2013. "Disagreement and Evidential Attenuation." *Noûs* 47 (4): 767–94. https://doi.org/10.1111/nous.12050.

———. 2014a. "Higher-Order Evidence and the Limits of Defeat." *Philosophy and Phenomenological Research* 88 (2): 314–45. https://doi.org/10.1111/phpr.12090.

———. 2014b. "The Dogmastism Puzzle." *Australasian Journal of Philosophy* 92 (3): 417–32. https://doi.org/10.1080/00048402.2013.834949.

Lee, Matthew. 2013. "Conciliationism Without Uniqueness." *Grazer Philosophische Studien* 88 (1): 161–88.

Levinstein, Ben. 2013. "Accuracy as Epistemic Utility." PhD thesis, Rutgers University.

Levy, Neil. 2005. "The Good, the Bad and the Blameworthy." *Journal of Ethics and Social Philosophy* 1 (2): 1–16. https://doi.org/10.26556/jesp.v1i2.6.

———. 2009. "Culpable Ignorance and Moral Responsibility: A Reply to Fitz-Patrick." *Ethics* 119 (4): 729–41. https://doi.org/10.1086/605018.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell Publishers.

———. 1978. "Truth in Fiction." *American Philosophical Quarterly* 15 (1): 37–46.

———. 1979. "Attitudes *de Dicto* and *de Se*." *Philosophical Review* 88 (4): 513–43. https://doi.org/10.2307/2184843.

———. 1982. "Logic for Equivocators." *Noûs* 16 (3): 431–41. https://doi.org/10.2307/2216219.

———. 1988. "Desire as Belief." *Mind* 97 (387): 323–32. https://doi.org/10.1093/mind/XCVII.387.323.

———. 1994. "Reduction of Mind." In *A Companion to the Philosophy of Mind*, edited by Samuel Guttenplan, 412–31. Oxford: Blackwell. https://doi.org/10.1017/CBO9780511625343.019.

———. 1996a. “Desire as Belief II.” *Mind* 105 (418): 303–13. https://doi.org/10.1093/mind/105.418.303.

———. 1996b. “Elusive Knowledge.” *Australasian Journal of Philosophy* 74 (4): 549–67. https://doi.org/10.1080/00048409612347521.

Lillehammer, Hallvard. 1997. “Smith on Moral Fetishism.” *Analysis* 57 (3): 187–95. https://doi.org/10.1111/1467-8284.00073.

Linton, Marisa. 2013. *Choosing Terror: Virtue, Friendship, and Authenticity in the French Revolution*. Oxford: Oxford University Press.

Lipsey, R. G., and Kelvin Lancaster. 1956-1957. “The General Theory of Second Best.” *Review of Economic Studies* 24 (1): 11–32. https://doi.org/10.2307/2296233.

Littlejohn, Clayton. 2012. *Justification and the Truth-Connection*. Cambridge: Cambridge University Press.

———. 2018. “Stop Making Sense? On a Puzzle about Rationality.” *Philosophy and Phenomenological Research* 96 (2): 257–72. https://doi.org/10.1111/phpr.12271.

Lockhart, Ted. 2000. *Moral Uncertainty and Its Consequences*. Oxford University Press.

Loewer, Barry, and Robert Laddaga. 1985. “Destroying the Consensus.” *Synthese* 62 (1): 79–95. https://doi.org/10.1007/BF00485388.

Lord, Errol. 2014. “From Independence to Conciliationism: An Obituary.” *Australasian Journal of Philosophy* 92 (2): 365–77. https://doi.org/10.1080/00048402.2013.829506.

MacAskill, William. 2014. “Normative Uncertainty.” PhD thesis, Oxford University.

———. 2016. “Normative Uncertainty as a Voting Problem.” *Mind* 125 (500): 967–1004. https://doi.org/10.1093/mind/fzv169.

Machuca, Diego E. 2013. “A Neo-Pyrrhonian Approach to the Epistemology of Disagreement.” In *Disagreement and Skepticism*, edited by Diego E. Machuca, 66–89. New York: Routledge.

Mahajan, Sanjoy. 2010. *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving*. Second. Cambridge, MA: MIT Press.

Maher, Patrick. 1997. “Depragmatized Dutch Book Arguments.” *Philosophy of Science* 64: 291–305. https://doi.org/10.1086/392552.

Maitra, Ishani, and Brian Weatherson. 2010. “Assertion, Knowledge and Action.” *Philosophical Studies* 149 (1): 99–118. https://doi.org/10.1007/s11098-010-9542-z.

Markovits, Julia. 2010. “Acting for the Right Reasons.” *Philosophical Review* 119 (2): 201–42. https://doi.org/10.1215/00318108-2009-037.

———. 2014. *Moral Reason*. Oxford: Oxford University Press.

Mason, Elinor. 2015. "Moral Ignorance and Blameworthiness." *Philosophical Studies* 172 (11): 3037–57. https://doi.org/10.1007/s11098-015-0456-7.

Matheson, Jonathan. 2009. "Conciliatory Views of Disagreement and Higher-Order Evidence." *Episteme* 6 (3): 269–79. https://doi.org/10.3366/E1742360009000707.

McKie, John, Peter Singer, Helga Kuhse, and Jeff Richardson. 1998. *The Allocation of Health Care Resources: An Ethical Evaluation of the 'QALY' Approach*. Aldergate: Ashgate.

McPhee, Peter. 2012. *Robespierre: A Revolutionary Life*. New Haven: Yale University Press.

Mogensen, Andreas L. 2016. "Contingency Anxiety and the Epistemology of Disagreement." *Pacific Philosophical Quarterly* 97 (4): 590–611. https://doi.org/10.1111/papq.12099.

Moller, D. 2011. "Abortion and Moral Risk." *Philosophy* 86 (3): 425–43. https://doi.org/10.1017/S0031819111000222.

Moody-Adams, Michelle M. 1994. "Culture, Responsibility and Affected Ignorance." *Ethics* 104 (2): 291–309. https://doi.org/10.1086/293601.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Morison, Benjamin. 2014. "Sextus Empiricus." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2014. Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/archives/spr2014/entries/sextus-empiricus/.

Moss, Sarah. 2011. "Scoring Rules and Epistemic Compromise." *Mind* 120 (480): 1053–69. https://doi.org/10.1093/mind/fzs007.

———. 2012. "Updating as Communication." *Philosophy and Phenomenological Research* 85 (2): 225–48. https://doi.org/10.1111/j.1933-1592.2011.00572.x.

———. 2015. "Time-Slice Epistemology and Action Under Indeterminacy." *Oxford Studies in Epistemology* 5: 172–94. https://doi.org/10.1093/acprof:oso/9780198722762.003.0006.

Nagel, Jennifer. 2013. "Defending the Evidential Value of Epistemic Intuitions: A Reply to Stich." *Philosophy and Phenomenological Research* 86 (1): 179–99. https://doi.org/10.1111/phpr.12008.

———. 2014. *Knowledge: A Very Short Introduction*. Oxford: Oxford University Press.

Nissan-Rozen, Ittay. 2015. "Against Moral Hedging." *Economics and Philosophy* 31 (3): 349–69. https://doi.org/10.1017/S0266267115000206.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In

Essays in Honor of Carl g. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday, edited by Nicholas Rescher, 114–46. Riedel: Springer.

———. 1981. *Philosophical Explorations*. Cambridge, MA: Harvard University Press.

———. 1994. *The Nature of Rationality*. Princeton: Princeton University Press.

Owens, David. 2002. "Epistemic Akrasia." *The Monist* 85 (3): 381–97.

Palmer, R. R. 1941. *Twelve Who Ruled*. Princeton, NJ: Princeton University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Peels, Rik. 2010. "What Is Ignorance?" *Philosophia* 38 (1): 57–67. https://doi.org/10.1007/s11406-009-9202-8.

Pettit, Philip, and Robert Sugden. 1989. "The Backward Induction Paradox." *Journal of Philosophy* 86 (4): 169–82. https://doi.org/10.2307/2026960.

Polymath, D. H. J. 2014. "New Equidistribution Estimates of Zhang Type, and Bounded Gaps Between Primes." http://arxiv.org/abs/1402.0811.

Price, Huw. 1989. "Defending Desire-as-Belief." *Mind* 98 (389): 119–27. https://doi.org/10.1093/mind/XCVIII.389.119.

Pryor, James. 2000. "The Sceptic and the Dogmatist." *Noûs* 34 (4): 517–49. https://doi.org/10.1111/0029-4624.00277.

———. 2004. "What's Wrong with Moore's Argument?" *Philosophical Issues* 14 (1): 349–78. https://doi.org/10.1111/j.1533-6077.2004.00034.x.

Quiggin, John. 1982. "A Theory of Anticipated Utility." *Journal of Economic Behavior & Organization* 3 (4): 323–43. https://doi.org/10.1016/0167-2681(82)90008-7.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13 (2): 134–71.

Regan, Donald. 1980. *Utilitarianism and Cooperation*. Oxford: Oxford University Press.

Reichenbach, Hans. 1956. *The Direction of Time*. Berkeley: University of California Press.

Ribeiro, Brian. 2011. "Epistemic Akrasia." *International Journal for the Study of Scepticism* 1: 18–25. https://doi.org/10.1163/221057011X554151.

Rosati, Connie S. 2016. "Moral Motivation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/archives/win2016/entries/moral-motivation/.

Rosen, Gideon. 2003. "Culpability and Ignorance." *Proceedings of the Aristotelian Society* 103 (1): 61–84. https://doi.org/10.1111/j.0066-7372.2003.00064.x.

———. 2004. "Skepticism about Moral Responsibility." *Philosophical Perspectives* 18 (1): 295–313. https://doi.org/10.1111/j.1520-8583.2004.00030.x.

———. 2008. "Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility." *Journal of Philosophy* 105 (10): 591–610. https://doi.org/10.5840/jphil20081051023.

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics* 116 (4): 742–68. https://doi.org/10.1086/505234.

Ross, Jacob, and Mark Schroeder. 2014. "Belief, Credence, and Pragmatic Encroachment." *Philosophy and Phenomenological Research* 88 (2): 259–88. https://doi.org/10.1111/j.1933-1592.2011.00552.x.

Rotondo, Andrew. 2013. "Undermining, Circularity, and Disagreement." *Synthese* 190 (3): 563–84. https://doi.org/10.1007/s11229-011-0050-2.

Russell, Bertrand. 1912/1997. *The Problems of Philosophy*. Oxford: Oxford University Press.

Russell, Jeffrey Sanford, and John Hawthorne. 2016. "General Dynamic Triviality Theorems." *Philosophical Review* 125 (3): 307–39. https://doi.org/10.1215/00318108-3516936.

Sartre, Jean-Paul. 1946/2007. "Existentialism Is a Humanism." In *Existentialism Is a Humanism*, translated by Annie Cohen-Solal, 17–72. New Haven: Yale University Press.

Schechter, Josh. 2013. "Rational Self-Doubt and the Failure of Closure." *Philosophical Studies* 163 (2): 429–52. https://doi.org/10.1007/s11098-011-9823-1.

Schoenfield, Miriam. 2014. "Permission to Believe: Why Permissivism Is True and What It Tells Us about Irrelevant Influences on Belief." *Noûs* 48: 193–218. https://doi.org/10.1111/nous.12006.

———. 2015. "A Dilemma for Calibrationism." *Philosophy and Phenomenological Research* 91 (2): 425–55. https://doi.org/10.1111/phpr.12125.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review* 117 (2): 245–73. https://doi.org/10.1215/00318108-2007-037.

———. 2011. "Self-Ignorance." In *Consciousness and the Self: New Essays*, edited by JeeLoo Liu and John Perry, 184–97. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511732355.009.

Scott, Rose M., and Renée Baillargeon. 2013. "Do Infants Really Expect Others to Act Efficiently? A Critical Test of the Rationality Principle." *Psychological Science* 24 (4): 466–74. https://doi.org/10.1177/0956797612457395.

Scurr, Ruth. 2006. *Fatal Purity: Robespierre and the French Revolution*. London: Chatto & Windus.

Sepielli, Andrew. 2009. "What to Do When You Don't Know What to Do." *Ox-*

ford Studies in Metaethics 4: 5–28.

Sharot, Tali. 2012. *The Optimism Bias: Why We're Wired to Look at the Bright Side*. London: Constable; Robinson.

Sidgwick, Henry. 1874. *The Methods of Ethics*. London: Macmillan.

Singer, Peter. 1972. "Famine, Affluence and Morality." *Philosophy and Public Affairs* 1 (3): 229–43.

Slote, Michael. 1992. *From Morality to Virtue*. Oxford: Oxford University Press.

Smart, J. J. C. 1961. *An Outline of a System of Utilitarian Ethics*. Melbourne: University of Melbourne Press.

Smith, Angela M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115 (2): 236–71. https://doi.org/10.1086/426957.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

———. 1996. "The Argument for Internalism: Reply to Miller." *Analysis* 56 (3): 175–84. https://doi.org/10.1111/j.0003-2638.1996.00175.x.

———. 2006. "Moore on the Right, the Good, and Uncertainty." In *Metaethics After Moore*, edited by Terrence Horgan and Mark Timmons, 133–48. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199269914.003.0007.

———. 2009. "Consequentialism and the Nearest and Dearest Objection." In *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson*, edited by Ian Ravenscroft, 237–66. Oxford: Oxford. https://doi.org/10.1093/acprof:oso/9780199267989.003.0011.

Smithies, Declan. 2012. "Moore's Paradox and the Accessibility of Justification." *Philosophy and Phenomenological Research* 85 (2): 273–300. https://doi.org/10.1111/j.1933-1592.2011.00506.x.

Srinivasan, Amia. 2015a. "Are We Luminous?" *Philosophy and Phenomenological Research* 90 (2): 294–319. https://doi.org/10.1111/phpr.12067.

———. 2015b. "Normativity Without Cartesian Privilege." *Philosophical Issues* 25: 273–99. https://doi.org/10.1111/phis.12059.

Staffel, Julia. 2015. "Disagreement and Epistemic Utility-Based Compromise." *Journal of Philosophical Logic* 44 (3): 273–86. https://doi.org/10.1007/s10992-014-9318-6.

Stalnaker, Robert. 1998. "Belief Revision in Games: Forward and Backward Induction." *Mathematical Social Sciences* 36 (1): 31–56. https://doi.org/10.1016/S0165-4896(98)00007-9.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford University Press.

Steup, Matthias. 2013. "Is Epistemic Circularity Bad?" *Res Philosophica* 90 (2): 215–35. https://doi.org/http:// dx.doi.org/ 10.11612/ resphil.2013.90.2.8.

Stoppard, Tom. 1967/1994. *Rosencrantz and Guildernstern Are Dead*. New York: Grove Press.

Strawson, P. F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.

Svavarsdóttir, Sigrún. 1999. "Moral Cognition and Motivation." *Philosophical Review* 108 (2): 161–219. https://doi.org/10.2307/2998300.

Tarsney, Christian. 2017. "Rationality and Moral Risk: A Moderate Defense of Hedging." PhD thesis, University of Maryland, College Park.

Titelbaum, Michael. 2014. *Quitting Certainties: A Bayesian Framework for Modeling Degrees of Belief*. Oxford: Oxford.

———. 2015. "Rationality's Fixed Point (or: In Defence of Right Reason)." *Oxford Studies in Epistemology* 5: 253–94. https://doi.org/10.1093/acprof:oso/9780198722762.003.0009.

———. 2016. "Self-Locating Credences." In *Oxford Handbook of Probability and Philosophy*, edited by Alan Hájek and Christopher Hitchcock, 666–80. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199607617.013.34.

Treynor, Jack L. 1987. "Market Efficiency and the Bean Jar Experiment." *Financial Analysts Journal* 43 (3): 50–53. https://doi.org/10.2469/faj.v43.n3.50.

Vargas, Manuel. 2005. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29 (1): 269–91. https://doi.org/10.1111/j.1475-4975.2005.00117.x.

Vavova, Katia. 2014. "Moral Disagreement and Moral Skepticism." *Philosophical Perspectives* 28 (1): 302–33. https://doi.org/10.1111/phpe.12049.

Vogel, Jonathan. 1990. "Cartesian Skepticism and Inference to the Best Explanation." *Journal of Philosophy* 87 (11): 658–66. https://doi.org/10.5840/jphil1990871123.

———. 2000. "Reliabilism Leveled." *Journal of Philosophy* 97 (11): 602–23. https://doi.org/10.2307/2678454.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–48.

Weatherson, Brian. 1999. "Begging the Question and Bayesians." *Studies in the History and Philosophy of Science Part A* 30: 687–97. https://doi.org/10.1016/S0039-3681(99)00020-5.

———. 2003. "Are You a Sim?" *Philosophical Quarterly* 53 (212): 425–31. https://doi.org/10.1111/1467-9213.00323.

———. 2004. "Luminous Margins." *Australasian Journal of Philosophy* 82 (3): 373–83. https://doi.org/10.1080/713659874.

———. 2005. "Scepticism, Rationalism and Externalism." *Oxford Studies in Epistemology* 1: 311–31.

———. 2012. "Knowledge, Bets and Interests." In *Knowledge Ascriptions*, edited by Jessica Brown and Mikkel Gerken, 75–103. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199693702.003.0004.

———. 2013. "The Role of Naturalness in Lewis's Theory of Meaning." *Journal for the History of Analytical Philosophy* 1 (10): 1–19. https://doi.org/10.4148/jhap.v1i10.1620.

———. 2014a. "Games, Beliefs and Credences." *Philosophy and Phenomenological Research* 92 (2): 209–36. https://doi.org/10.1111/phpr.12088.

———. 2014b. "Probability and Scepticism." In *Scepticism and Perceptual Justification*, edited by Dylan Dodd and Elia Zardini, 71–86. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199658343.003.0004.

———. 2015. "Memory, Belief and Time." *Canadian Journal of Philosophy* 45 (5-6): 692–715. https://doi.org/10.1080/00455091.2015.1125250.

Wedgwood, Ralph. 2012. "Justified Inference." *Synthese* 189 (2): 273–95. https://doi.org/10.1007/s11229-011-0012-8.

Weirich, Paul. 2016. "Causal Decision Theory." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/archives/win2016/entries/decision-causal/.

Weisberg, Jonathan. 2010. "Bootstrapping in General." *Philosophy and Phenomenological Research* 81 (3): 525–48. https://doi.org/10.1111/j.1933-1592.2010.00448.x.

White, Roger. 2005. "Epistemic Permissiveness." *Philosophical Perspectives* 19 (1): 445–59. https://doi.org/10.1111/j.1520-8583.2005.00069.x.

———. 2006. "Problems for Dogmatism." *Philosophical Studies* 131: 525–57. https://doi.org/10.1007/s11098-004-7487-9.

———. 2009. "On Treating Oneself and Others as Thermometers." *Episteme* 6 (3): 233–50. https://doi.org/10.3366/E1742360009000689.

Williams, Bernard. 1981. "Persons, Character, and Morality." In *Moral Luck*, 1–19. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139165860.002.

———. 1995. "Saint-Just's Illusion." In *Making Sense of Humanity and Other Philosophical Essays*, 135–50. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511621246.013.

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

———. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

———. 2011. "Improbable Knowing." In *Evidentialism and Its Discontents*, edited by T. Dougherty, 147–64. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199563500.003.0010.

———. 2014. "Very Improbable Knowing." *Erkenntnis* 79 (5): 971–99. https://doi.org/10.1007/s10670-013-9590-9.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. London: Macmillan.

Wolf, Susan. 1987. "Sanity and the Metaphysics of Responsibility." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand David Schoeman, 46–62. Cambridge University Press. https://doi.org/10.1017/cbo9780511625411.003.

Worsnip, Alex. 2014. "Disagreement about Disagreement? What Disagreement about Disagreement?" *Philosophers' Imprint* 14 (18): 1–20. http://hdl.handle.net/2027/spo.3521354.0014.018.

Wright, Crispin. 2000. "Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof." *Philosophical Issues* 10: 140–63. https://doi.org/10.1111/j.1758-2237.2000.tb00018.x.

———. 2002. "(Anti-)sceptics Simple and Subtle: G.e. Moore and John McDowell." *Philosophy and Phenomenological Research* 65 (2): 330–48. https://doi.org/10.1111/j.1933-1592.2002.tb00205.x.

Zimmerman, Michael J. 2008. *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge: Cambridge University Press.