The Scope and Limits of Debunking Arguments in Ethics

Shang Long Yeo
June 2020

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University.

Declaration

This thesis is solely the work of its author. No part of it has previously been submitted for any degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this thesis, and all sources used, have been duly acknowledged.

Word Count: 60512 words

Shang Long Yeo,

22 Jun 2020

Acknowledgements

Doing a PhD at the ANU School of Philosophy has been the academic adventure of a lifetime, one that has exceeded all expectations. I am humbled by the friendship, love, and support I've gotten here, and Canberra will always have a special place in my heart.

Thanks first to Christian Barry, the chair of my panel, for being a superb supervisor and a supportive friend over these years. Thank you for your insight and your patience, for gently challenging me to do better, and for your hard work in making this thesis come to life. Thanks also to my awesome supervisory panel: Thanks to Al Hájek, for your enthusiasm and positivity, for your numerous comments, and for always being up to chat at drinks and beyond. Thanks to Kim Sterelny, for your encouragement and expert guidance, and for commiserating with the perfectly placed expletive when life called for it. Thanks to Katie Steele, for your thoughtful and constructive advice, and for your constant support and good humour.

Many thanks to everyone at the ANU – students, faculty, visitors – for such a philosophically enriching time. Thanks especially to the following for conversations and comments about my work: Geoff Brennan, Rachael Brown, Devon Cass, Juan Comesaña, Bronwyn Finnigan, Jesse Hambly, Toby Handfield, Josef Holden, Jessica Isserow, Joshua Knobe, Ole Koksvik, Ten-Herng Lai, Seth Lazar, Chris Lernpass, Matt Lindauer, Edouard Machery, Kirsten Mann, Stephen Mann, Don Nordblom, Philip Pettit, Oliver Rawle, Vince Redhouse, Thomas Schmidt, Walter Sinnott-Armstrong, Toby Solomon, Daniel Stoljar, Jeremy Strasser, Hezki Symonds, Joshua Thong, Lachlan Umbers, Lachlan Walmsley, James Willoughby, and Brandon Yip. Thanks also to Abhishek Mishra, Neil Sinhababu, and Weng Hong Tang.

I'm also grateful to audiences at: various Philsoc seminars, Kioloa conferences, and MSPT Graduate Workshops at the ANU, the 2016, 2017 and 2018 Australasian Postgraduate Philosophy Conferences, the 2018 and 2019 iterations of the ANU-Humboldt-Princeton Summer Institute, the 2019 Australasian Association of Philosophy Conference, and the International Conference on Moral Epistemology at the Australian Catholic University

Melbourne. Thanks also to anonymous reviewers from various journals for their helpful comments.

I'm very grateful for the funding I've received, particularly from the Australian Government. This research is supported by an Australian Government Research Training Program (RTP) Scholarship. I'm also thankful for additional support from the National University of Singapore's Overseas Graduate Scholarship.

A big thanks to my friends for supporting and encouraging me in all the right ways. I would write about how you've changed my life, but that will require a couple hundred more pages. So I'll just say: thank you for being there for me. Thanks especially to Lenore, Kirsten, Jesse, Josef, Hezki, Toby, James, Ste, Devon, Jeremy, Don, Bruno, Heather, Brenna, Szymon, Natha, Abhishek. And thanks to my family, for supporting me and letting me chase my dreams.

Abstract

Debunking arguments use empirical evidence about our moral beliefs – in particular, about their causal origins, or about how they depend on various causes – in order to reach an epistemic conclusion about the trustworthiness of such beliefs. In this thesis, I investigate the scope and limits of debunking arguments, and their implications for what we should believe about morality. I argue that debunking arguments can in principle work – they are based on plausible epistemic premises, and at least some of them avoid putative problems concerning regress and redundancy. However, I also argue that some debunking arguments fall short because they are insufficiently supported by the empirical evidence. By considering different objections, analyses, and a case study, I explore the conditions for a successful debunking argument.

Chapter 1 starts by providing an overview of debunking arguments – their structure, their variations, and common objections to them. Chapter 2 defends debunking arguments against counterexamples to their epistemic premises – counterexamples which, if effective, will show that debunking arguments cannot work in principle. I argue against the use of these counterexamples, and contend that we cannot merely deny the epistemic premises of a debunking argument. Chapter 3 defends debunking arguments against three further objections concerning how such arguments fit into the web of beliefs. The regress objection contends that debunking arguments make assumptions that commit us to a problematic regress. The findings redundancy objection contends that the empirical findings are unnecessary in a debunking argument. The argument redundancy objection alleges that debunking arguments assume what they set out to prove.

Even if debunking arguments are in principle viable, some of them still fail because of poor empirical support. Chapter 4 argues that some global, evolutionary debunking arguments fall short – roughly because we are poorly placed to observe, intervene on, and predict what would happen from different evolutionary causes. In contrast, experimental evidence from moral psychology and behavioural economics will be better placed to support a debunking

argument. Chapter 5 considers a novel debunking argument based on evidence of this kind – which shows how we overweight low probabilities in our decision-making, and underweight moderate to high probabilities. I explore debunking and vindicating arguments concerning our intuitions about risky aid.

Chapter 6 proposes a Bayesian analysis of debunking arguments, which can guide us in revising our beliefs in response to debunking. I highlight further conditions for debunking to work, and propose a quantitative method for integrating different kinds of evidence in order to arrive at a rational credence about morality in light of debunking.

Table of Contents

Introduction

You have some moral beliefs – about whether we're morally permitted to privilege ourselves over others, about whether we're permitted to eat factory-farmed meat, about when (if ever) we're required to aid others, and so on. These beliefs came from somewhere – they had *origins*. Maybe you arrived at such beliefs through arguments and reflection, or from your parents and society, or via some innate biological mechanism, or through some combination of these and other processes. Learning about the origins might also shed light on how your holding of these beliefs depends on various causes – you might realise, for instance, that you would have different moral beliefs (or no moral beliefs at all) if you had been born to a different society, or descended from ancestors who were subject to different evolutionary pressures, or encountered a different set of thought experiments. Alternatively, the origins might also reveal how your beliefs are unresponsive to the facts – you might realise that had the facts been changed, you would still have held the same moral beliefs.

Could we learn anything about the origins of our moral beliefs,[1] or about how our holding of these beliefs depends on various causes, that would lead us to conclude that such beliefs are untrustworthy? Some philosophers think so. They argue that once we discover the origins of some, or all, of our moral beliefs, we should think that such beliefs are epistemically flawed, and hence that they're untrustworthy. These philosophers endorse what's called a *debunking argument* in the philosophical literature.[2]

---

[1] When I speak of moral beliefs in this thesis, I mean to refer only to positive moral beliefs. A positive moral belief is one that attributes a moral property to an action or entity – this includes, for instance, the belief that "Eating factory-farmed meat has the property of being morally forbidden" and the belief that "Ginny has the property of being courageous". This definition excludes beliefs which *deny* the instantiation of moral properties, such as the belief that "Breaking a promise *does not* have the property of being morally forbidden".

[2] For evolutionary debunking arguments, see Joyce (2006, 2016), Street (2006), Ruse and Wilson (1986), Fraser (2014), Handfield (2016), Morton (2016), Bogardus (2016), de Lazari-Radek and Singer (2012). For debunking arguments that rely on evidence from experimental psychology, see Horowitz (1998), Greene (2008), Alfano (2011), Liao et al. (2012), O'Neill (2015). A significant part of the debunking literature speaks about our *normative* beliefs more broadly – but I'll only focus on how the arguments and objections apply to our moral beliefs.

Here are two influential examples of such arguments: first, Joyce (2006) argues that once we recognise the contingent evolutionary origins of our moral concepts, all our moral beliefs are undermined. Second, Liao et al. (2012) attempt to undermine moral intuitions[3] about the Loop case: where a trolley is headed toward five innocent people and will kill these five, but could be diverted to a side track to kill one innocent person. This side track loops back to the main track with five innocent people – so that if there wasn't a person on the side track, the trolley would loop back and still kill the five (see below).

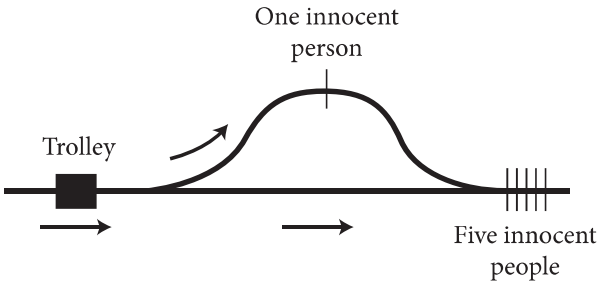


FIG 1.1 THE LOOP CASE

Liao et al. find that subjects' moral intuitions about the Loop case varies with the case's order of presentation. They conclude that such intuitions shouldn't be trusted – and, presumably, neither should any beliefs based on such intuitions.

Despite their variety, I believe debunking arguments can be understood in terms of the common schema below:

> (*Empirical Evidence* about the causal origins of the target moral beliefs, and how your holding of such beliefs depends on various causes)
>
> (*Evidence Reveals Flaw*) If *Empirical Evidence*, then the target moral beliefs have some epistemic flaw F (e.g. they are too contingent, they are unresponsive to the facts, etc).
>
> (*Epistemic Premise*) If a belief has flaw F, then it is epistemically undermined (e.g. it is unjustified, it doesn't count as knowledge, it is likely false, etc).

[3] In this thesis, I will only assume that the intuition that p is not identical to a belief that p. Otherwise I remain neutral about the exact nature of intuitions – they might be dispositions to believe, intellectual seemings, or some other sui generis mental state (Pust, 2019, sec. 1).

(*Epistemic Conclusion*) Therefore, the target moral beliefs are epistemically undermined. (from *Empirical Evidence, Evidence Reveals Flaw* and *Epistemic Premise*)

Empirical Evidence states findings about the target moral beliefs – usually findings from experimental and evolutionary psychology, or from related fields like behavioural economics. Such findings allegedly reveal an epistemic flaw of the target beliefs (*Evidence Reveals Flaw*) – for instance, that these beliefs are too contingent, or that they are unresponsive to the facts. In light of this flaw, the target beliefs are epistemically undermined (through *Epistemic Premise*). Together, these premises generate a skeptical threat against the target moral beliefs – these target beliefs are undermined, or, more generally, we should trust them less than we used to (*Epistemic Conclusion*).

1. Why You Should Care

Why care about debunking arguments? First, they purport to bring new kinds of evidence – about the causal origins of our moral beliefs – to bear in assessing the epistemic status of such beliefs. In doing so, such arguments present a new way of evaluating our moral beliefs – over and above more traditional methods of trying to achieve coherence with other moral beliefs, or discarding moral beliefs that were based on false empirical presuppositions.

Secondly, debunking arguments have been invoked to support a variety of positions in metaethics and normative ethics. Debunking arguments might undermine the beliefs or intuitions[4] that support the existence of moral facts, thus strengthening the overall case for claiming that such facts don't exist (Joyce, 2009, pp. 217–220). Alternatively, such arguments might help a moral skeptic prove the epistemological thesis that we can't justifiably believe or know anything about the moral facts, even while conceding the metaphysical thesis that such facts might exist (Bogardus, 2016, p. 637). Debunking arguments might also undermine

[4] In my framework, I will take debunking/undermining an intuition to be equivalent to undermining a belief that was based on this intuition.

some, but not all, of our moral beliefs – and thus help us rationally adjudicate disagreements and clashes of intuitions about various normative questions (Singer, 2005; Huemer, 2008), for instance about whether or not it is typically wrong to consume animal products (McPherson, 2014).

Thirdly, debunking arguments purport to require only minimal assumptions in order to work.[5] Such arguments might thus provide a relatively theory-neutral way of evaluating the inputs to moral theorising – one that does not beg the question against either side of many moral debates. Here are two initial assumptions that I will just take for granted in the rest of this thesis: a) that our moral judgments are apt to be true or false (the plausible claim of moral cognitivism), b) that a simple subjectivism (which holds that the moral fact that p obtains iff I believe that p) is false.[6] As will become clear in later chapters, however, specific debunking arguments will require further assumptions about what the moral facts could be like.

These three features of debunking arguments – their purportedly minimal assumptions, their wide range of application, and the novelty of the evidence that they draw on – make such arguments a potentially powerful tool for moral theorising. Thus there is a pressing need to understand how such arguments work, and what conclusions are licensed by such arguments.

2. The Thesis

In this thesis, I investigate the structure and content of debunking arguments, and their implications for what we should believe about morality. I argue that debunking arguments can in principle work – such arguments are based on plausible epistemic premises, and at least some of them avoid putative problems concerning regress and redundancy. However,

[5] For instance, Joyce (2016, n. 8) seems to think that his debunking argument only requires the assumption of moral cognitivism. Unfortunately, many existing debunking arguments (particularly evolutionary ones) in fact make strong metaphysical assumptions about the nonexistence of the moral facts and the impossibility of reducing the moral to the non-moral (Das, 2016), or must at least assume something about what the moral facts could be like (Vavova, 2014, pp. 92–93). I believe that some metaphysical assumptions will inevitably be needed in debunking, but I also believe that these could be quite minimal, and I will flag them when required.

[6] If this subjectivist theory is true, then our moral beliefs are guaranteed to be true, and they will not be vulnerable to a debunking argument. Thanks here to Jessica Isserow.

some of these arguments still fall short because of poor empirical support. By considering different objections, analyses, and a case study, I investigate the conditions for a successful debunking argument. Roughly, I believe that a debunking argument will be successful only if a) it bears on the probability that the target moral beliefs are true (rather than on their status as justified or knowledge), b) it relies on assumptions that are more likely to be true than the target beliefs, c) it does not assume that the target beliefs are false, d) it relies on empirical evidence which reveals something that hasn't thus far been recognised from the armchair alone, and e) this empirical evidence can establish the relevant claims about how the target beliefs depend on various causes.

I believe my investigation is distinctive in several ways. First, I distinguish between the different epistemic standards involved in debunking and argue explicitly for one of them. Second, I examine how debunking arguments fit into the broader web of beliefs, and the potential issues that might arise from this. Third, I pay close attention to how specific empirical findings could generate epistemic implications. Finally, I recognise that the impact of debunking can come in degrees, and I provide a method for quantifying this impact.

I'll now outline the chapters to come. Chapter 1 reviews debunking arguments – their structure, their variations, and some common objections to them. Chapter 2 defends debunking arguments against counterexamples to their *Epistemic Premise* – which, recall, undermines the target beliefs on the basis of their epistemic flaws. If such counterexamples are effective, then debunking arguments cannot work in principle. I argue instead that these counterexamples are either based on inappropriate epistemic standards, or on misunderstandings about what counts as a flawed belief. I also argue more generally that we cannot rest content with mere denials of *Epistemic Premise*, on pain of being overly permissive with clearly problematic beliefs. Chapter 3 defends debunking arguments against three further objections, which all pertain to the structure of support in moral epistemology. The *regress* objection contends that debunking arguments require moral assumptions which commit us to a problematic regress, and that this regress disables the debunking conclusion. The *findings redundancy* objection contends that the empirical findings are unnecessary in a

debunking argument, and that a debunking argument is really just armchair reasoning in disguise. The *argument redundancy* objection alleges that debunking arguments assume what they set out to prove, such that their undermining effect is evidentially redundant.

Even if debunking arguments can be successful, some of them fall short because they aren't sufficiently supported by the empirical evidence. In Chapter 4, I argue that some global, evolutionary debunking arguments are unsuccessful, roughly because we are poorly placed to observe, intervene on, and predict what would happen from different evolutionary causes. In particular, I argue that *Evidence Reveals Flaw* is implausible in these arguments – the empirical evidence is unable to establish that our moral beliefs are flawed in the relevant ways. In contrast, I shall argue that experimental evidence from moral psychology and behavioural economics would more likely support a debunking argument. Chapter 5 considers a novel debunking argument based on evidence of this kind – behavioural economists have conducted experiments showing how we overweight low probabilities in our decision-making, and underweight moderate to high probabilities. I use our intuitions about risky aiding as a case study, exploring possible debunking and vindicating arguments about them.

Suppose that a debunking argument works – how should we change our moral beliefs in response? In Chapter 6, I propose a Bayesian analysis of debunking arguments – which proceeds in the fine-grained framework of rational confidence, as opposed to the coarse-grained notion of justified binary belief assumed by most of the literature. I identify further conditions for debunking to work, and propose a quantitative method for integrating different kinds of evidence – about the types of epistemic flaws at play, about the different possible origins of our moral beliefs, and about the background normative assumptions that we're entitled to make – in order to arrive at a rational credence about morality, in light of debunking.

The resulting view of debunking arguments is nuanced – some of these arguments work, and some don't. In this thesis, I hope to explore when debunking arguments are successful and can contribute to moral theory. In other words, then, I hope to investigate the scope and limits of debunking arguments in ethics.

1 Debunking Arguments: An Overview

I'll start by giving an overview of debunking arguments. In this chapter, I outline some examples of such arguments, highlight their common structure, and illustrate how the different parts could vary from argument to argument. I then review some common objections. Finally, I reflect on further issues concerning debunking.

1.1 Anatomy of a Debunking Argument

To help anchor our discussion, recall two debunking arguments referred to in the introduction. First, Joyce (2006) argues that our tendency to use moral concepts – obligation, virtue, property, desert and so on – is highly contingent on how human evolutionary history has played out. Had our evolutionary history played out differently, we would have very different moral beliefs, or no moral beliefs at all. This, he thinks, should lead us to be agnostic about all our moral beliefs. Secondly, Liao et al. (2012) argue that our moral intuitions about the Loop case are epistemically problematic because they vary with the case's order of presentation. Our moral intuitions about the Loop case change depending on what case was seen before it – yet the moral facts remain the same, regardless of which case we saw first. This, you might think, should make us suspicious about our moral intuitions – and the corresponding moral beliefs – about the Loop case.

These two examples highlight how diverse debunking arguments can be. Some rely on evidence about the proximate causes of moral belief (that is, causes that occur within our lifetimes, such as the immediate psychological mechanisms that produce our moral intuitions or beliefs); others use evidence about more distant causes (often termed 'ultimate causes', which are historical causes outside of our lifetimes, like natural selection over many generations). Some target all our moral beliefs at one go (creating a global debunking argument); others only hope to undermine a limited subset (local debunking). Still, such arguments share a common structure, as outlined in the schema from the introduction:

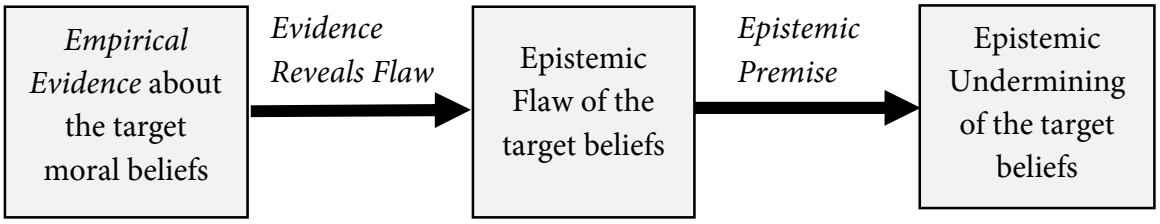
(*Empirical Evidence* about the causal origins of the target moral beliefs, and how your holding of such beliefs depends on various causes)

(*Evidence Reveals Flaw*) If *Empirical Evidence*, then the target moral beliefs have some epistemic flaw F.

(*Epistemic Premise*) If a belief has flaw F, then it is epistemically undermined.

(*Epistemic Conclusion*) Therefore, the target moral beliefs are epistemically undermined. (from *Empirical Evidence, Evidence Reveals Flaw* and *Epistemic Premise*)

This schema can equally well be put in terms of the flowchart below:



Think about what the boxes of the flowchart would be in Joyce's and Liao et al.'s arguments. Joyce's *Empirical Evidence* consists of findings that almost all human societies have moral concepts, and that children acquire the ability to reason with deontic conditionals from a young age (Joyce, 2006, pp. 134–136). He uses this to argue that the tendency to believe in moral reasons is an evolutionary adaptation – it facilitated cooperation among our ancestors, leading to their reproductive success. This tendency was then genetically inherited by us modern humans. This might show that our moral beliefs are unacceptably chancy or contingent (the Epistemic Flaw), because it could easily have been that our ancestors didn't face the same evolutionary pressures, which would have led to their (and our) not believing in

moral reasons at all.[1] Joyce then argues that learning about this contingency should lead us to be agnostic about all our moral beliefs (the Epistemic Undermining).[2]

Liao et al.'s *Empirical Evidence* consists of experimental findings about how subjects rated the moral permissibility of turning the trolley in the Loop case. They compared subjects' intuitions about the Loop case when it was the first case seen, versus when Loop was seen after the Standard case (where there is a side track with one person, but that track doesn't loop back), versus when Loop was seen after the Push case (where, instead of a side track, there is one person standing on a bridge over the track, and subjects are asked if it's permissible to push this one person over, in order to stop the trolley and save the five). Subjects responded to Loop differently, depending on whether they saw Loop first, or whether they saw it after Standard or Push. This might also show that moral intuitions about the Loop case are unacceptably chancy (the Epistemic Flaw), which leads to the conclusion that beliefs based on such intuitions are undermined (the Epistemic Undermining).

These examples give the general flavour of how a debunking argument works – although the different parts can vary significantly. I'll now focus in on each part of the argument, cataloguing some prominent variations. I'll work backwards with the boxes of the flowchart – starting first with the Epistemic Undermining.

1.1.1 The Epistemic Undermining

Debunking arguments conclude that the target beliefs are epistemically undermined, but they don't say that the target beliefs are *false*. An analogy will help here.[3] Say you're looking at a

[1] Alternatively, it might reveal a debunking explanation that doesn't presuppose the truth of our moral beliefs (Joyce, 2006, p. 184) – see the later section on explanation-based flaws.

[2] Alternatively, Joyce has also argued that these origins render our moral beliefs unjustified. I believe this is a different kind of epistemic undermining, however – see Chapter 2.

[3] The following is a familiar case from the literature on undercutting defeaters in epistemology, started by Pollock (1970) – see Sudduth (2019) for a comprehensive review. McGrath (2014, pp. 210–211), Lutz (2018), Kahane (2011, p. 106), Bogardus (2016, p. 637) all see the debunking argument as providing an undercutting defeater for our moral beliefs.

wall, and it appears red to you. You trust your perception of the red wall, so you go on to believe that the wall is red. I then inform you, however, that there is a red light shining on the wall. Upon learning this, it seems you should be less confident that what you actually saw was a red wall. Why? Because the wall could have been white, yet still appear red to you because of the red light shining on it. Upon learning about the red light, you should think that the wall's appearing red is an untrustworthy indication of the wall's actual colour. That's not to say that you should conclude that the wall is *not* red – rather, you just shouldn't give as much evidentiary weight to your perception, and should perhaps be agnostic about the wall's colour for now.

Debunking arguments contend that learning about the causes of the target beliefs is like learning about the red light shining on the wall. Just as how we should become distrustful of our belief that there is a red wall upon learning about the light, we should likewise be suspicious of the target moral beliefs upon learning about their origins.

This Epistemic Undermining of the target beliefs could come in several different variants. Debunking arguments could conclude that the target moral beliefs…

- are *unjustified*, or *pro tanto unjustified*;
- merely *do not count as knowledge*;
- are *likely false*, or are *more likely than before to be false*.[4]

In Chapter 2, I'll argue for one way of understanding the Epistemic Undermining in a debunking argument, thus setting the scope of our inquiry. For now, notice that whichever variant of undermining you adopt, the debunking argument's conclusion is about the epistemic status of the target moral beliefs – and not about the nature of the moral facts themselves. In other words, the conclusion of a debunking argument is epistemic, not metaphysical. This sets debunking arguments apart from older arguments which seek to establish that there are no moral facts, such as Mackie (1977) (Vavova, 2015, p. 105). (One

[4] For examples of each kind of epistemic undermining, see Joyce (2006, p. 181), Bogardus (2016, pp. 657–658), and Street (2006, p. 125) respectively.

complication: some proponents of debunking, like Street (2006), want to get metaphysical conclusions from premises about the origins of our moral beliefs. Still, they do so *through* an intermediate epistemic conclusion like the one I've been talking about – so I'll keep talking about the epistemic conclusion for now, and only return to the metaphysical extensions in section 4.)

Reducing the epistemic status of the target moral beliefs is a significant achievement. This doesn't mean, of course, that those beliefs couldn't have high epistemic status, all things considered. This status could be restored by other means that aren't debunked, such as through other independent arguments (Joyce, 2016, p. 125; Kahane, 2011, pp. 106–107).

You might also wonder *whose* beliefs are being undermined.[5] For the purposes of this thesis, I'll be examining the epistemic support that we *philosophers* – with all available information about the origins of our moral beliefs – have for holding the moral beliefs that we do. I leave open the question of whether the folk have trustworthy moral beliefs or not.

It's also worth distinguishing the undermining that would seem warranted in light of debunking arguments from other kinds of undermining. One is what White (2010, pp. 575–576) calls reassessment. Sometimes, the empirical evidence prompts us to think about the reasons we have for our moral beliefs. We might, as a result of this reassessment, realise that we don't have good reasons for our beliefs – and this undermines the support for such beliefs. But, as White points out, the empirical evidence isn't crucial here: even if we realize later that the evidence was misleading, we would still stick with the new set of beliefs (rather than revert to the old ones), since we still don't have good reasons for the old beliefs. I grant that reassessment can undermine the support for our moral beliefs – but since the empirical evidence is merely *triggering* the undermining rather than *justifying* it here, I will set aside reassessment for the purposes of this thesis.

The empirical evidence might also undermine the support supplied by the widespread agreement of others (Kahane, 2011, p. 110; Parfit, 1986, p. 186; White, 2010, p. 586). Suppose

[5] See Enoch (2010, p. 427), Nichols (2014, n. 11), and Isserow (2018, sec. 6) for some discussion of this.

I observe that everyone around me believes that p – and on that basis, I also start believing p. I grant that some empirical evidence might neutralize the support granted by everyone's believing that p – perhaps by showing that everyone's beliefs are not as independent of each other's as originally thought. But this is often not what the debunkers are aiming for – so I will also set aside this kind of epistemic impact, focusing instead on how the empirical evidence could undermine the support rendered by other kinds of evidence (like an individual's own intuitions or beliefs).

1.1.2 The Epistemic Flaw

The Epistemic Undermining of the target beliefs is brought about by revealing an Epistemic Flaw of such beliefs. Three kinds of flaws are discussed in the literature. *Process-based* flaws concern the process that created the target beliefs. *Explanation-based* flaws pertain to the plausibility of different explanations about the relationship between the target beliefs and the moral facts. *Disagreement-based* flaws relate to the presence of disagreement about the target beliefs. I'll now outline each type of flaw, and explain an important distinction that cross-cuts them.

- **Process-based flaws**

Any belief will have been produced by some belief-forming process – like reasoning, introspection, wishful thinking, and so on. Such processes are reliable to the extent that they produce true beliefs. A belief-forming process is said to be reliable if it produces a large proportion (at least greater than 50%) of true beliefs; otherwise it is unreliable. This reliability can be measured as the proportion of true beliefs produced by the process in the actual world (actual reliability), or as the proportion of true beliefs produced in the actual world and some possible worlds (counterfactual reliability) (Goldman & Beddor, 2016).

Many debunking arguments contend that our moral belief-forming processes are unreliable, either in the actual or counterfactual sense. Fraser (2014) argues, for instance, that evolutionary considerations show that our moral belief-forming processes are *actually*

unreliable – because such processes fail to meet various conditions pertaining to the environment in which the process was selected, the relative costs of accuracy versus error, the relative costs of different kinds of error, and the evolved function of the process.

Others argue instead that our moral belief-forming processes are *counterfactually* unreliable. This charge of counterfactual unreliability comes in two flavours. The debunkers could allege that our moral beliefs are unacceptably *chancy*, contingent, lucky, or accidental – given the process that produced these beliefs, we could easily have arrived at different moral beliefs, even while the moral facts remain the same. This is often interpreted as an argument that our moral beliefs are unsafe – roughly, a belief that p is safe iff we could not easily have falsely believed that p.[6] Joyce (2006) might provide an example of this challenge: he can be read as arguing that our moral beliefs are unacceptably contingent on how our specific evolutionary history has played out – that we could easily have followed a slightly different evolutionary trajectory, such that we would have different moral beliefs, or no moral beliefs at all. Alternatively, the debunkers might allege that our moral beliefs are unacceptably *inevitable* or rigid: our moral belief forming processes were such that we were bound to have the same moral beliefs, even when the moral facts are changed. This is normally interpreted as arguing that our moral beliefs are insensitive – roughly, a belief that p is sensitive iff had p been false, we wouldn't believe that p.[7] Ruse and Wilson (1986) exemplify this kind of challenge: they argue that even if there weren't any moral facts, we would still have believed that there were – because evolution would have forcefully inculcated moral beliefs in us anyway.

[6] Debunking arguments based on safety – or the related condition of adherence – are discussed by Joyce (2006, p. 181), Clarke-Doane (2016, pp. 28–29), Bogardus (2016, pp. 645–647), Handfield (2016, p. 68), Shafer-Landau (2012, pp. 17–18), O'Neill (2015), Srinivasan (2015, pp. 337–341), Lillehammer (2010). For general overviews of safety and adherence, see Rabinowitz (2018), Comesaña (2007, pp. 786–789) and Nozick (1981, pp. 172–196).
[7] Debunking based on sensitivity – or on the screening-off of evidence – are discussed by Joyce (2006, p. 181), Bogardus (2016, pp. 638–640), Clarke-Doane (2012), Shafer-Landau (2012, pp. 16–18), White (2010, pp. 580–581), Kahane (2011, p. 106), Schafer (2010), Morton (2016, p. 235,242-243), O'Neill (2015), Ruse and Wilson (1986, pp. 186–187), Srinivasan (2015, pp. 329–330), Wielenberg (2010, pp. 454–456), Sinclair (2018, sec. 5.1). For general introductions to sensitivity, see Comesaña (2007, pp. 782–786) and Nozick (1981, pp. 172–196).

- **Explanation-based flaws**

For any moral belief we hold, we will likely have causal explanations for that belief – that is, theories or claims describing how we were caused to hold that belief. The debunkers could undermine our moral beliefs by pointing to explanation-based flaws, which draw on the plausibility of such causal explanations. Some causal explanations of a belief don't presuppose or entail that the belief is true or highly likely to be true[8] – call these *debunking explanations* (Tersman, 2008, pp. 394–395). Other explanations of a belief do presuppose or entail that that belief is true or significantly likely to be true – call these *vindicating explanations*.

A debunker might argue that the target moral beliefs are undermined because we have a sufficiently good debunking explanation of them. Joyce (2006, p. 184), for instance, also argues that we have an empirically-supported theory of our moral beliefs that nowhere presupposes their truth – this, he thinks, undermines such beliefs. Others argue that the problem lies in not having a sufficiently good vindicating explanation of our moral beliefs.

A different kind of argument relies on a reliability challenge – a demand to explain how our moral beliefs track the moral facts. In particular, Enoch (2010, pp. 420–425) sees the debunking argument as a challenge to explain how our moral beliefs could be well-correlated with the moral facts.[9] I believe this challenge reduces to the demand to show that our moral beliefs are produced by a reliable process, either in the actual or counterfactual sense (see previous section) – because if our moral beliefs are well-correlated with the moral facts, these beliefs would be mostly true, which would mean that the process that produced such beliefs is reliable.

[8] There's probably no significant difference between an explanation that *presupposes* the truth of a belief as opposed to one that *entails* it, but I have some preference for characterizing the debunking explanation in terms of entailment, since it fits better with the project of finding out whether our moral beliefs are undermined or not.
[9] See also Schechter (2017) on the reliability challenge.

- **Disagreement-based flaws**

Finally, disagreement-based flaws draw on potential disagreement about the target beliefs. Debunking arguments in this vein allege that there is epistemically significant disagreement concerning the target beliefs, and that this undermines such beliefs. However, these debunking arguments don't usually point to actual agents who disagree with us about the target moral beliefs – since arguments from actual moral disagreement are already well worked-out. Instead, these arguments cite disagreement with nearby *counterfactual* agents – agents who could easily have existed, but didn't, usually for some epistemically arbitrary reason. For instance, Bogardus (2016, pp. 655–659) argues that the evolutionary findings reveal nearby counterfactual selves who disagree with us about moral matters. He argues that we can't ascertain whether we are in a better epistemic position than our counterfactual selves – we are in a position of epistemic symmetry with them – and this means that our moral beliefs don't count as knowledge.

1.1.3 Further Flaws, Interrelations, Issues, and a Distinction

There are two further flaws that I'll mention for completeness, but will set aside for most of this thesis. First, some argue that it would be a massive coincidence if our moral beliefs turned out to be true – since there is a multiplicity of ways the moral facts could be, and evolutionary pressures pushed us toward just one particular set of moral beliefs. Thus, it is highly unlikely that this set of moral beliefs is true.[10] Secondly, others might argue that our moral beliefs lack independent confirmation from non-moral sources – in order to argue that our moral beliefs are reliable, we need to appeal to our moral beliefs.[11] Debunking arguments which invoke these two flaws don't seem to rely on the empirical evidence, so I won't consider such flaws for the rest of this thesis.

[10] Street (2006, p. 122) suggests this line of argument at points.

[11] Berker (2014, pp. 248–250) argues that if the evolutionary debunking argument is read in this way, then it just presses a general skeptical worry. See the discussion of third-factor accounts below.

There are also interesting relationships between the different flaws – I'll confine myself to two examples here. First, when a belief becomes less chancy or contingent, the less it's going to be affected by different causes, and so the more likely that it's going to be unacceptably rigid. White (2010, p. 579) protests that this means it's 'damned if you do, damned if you don't' when it comes to our moral beliefs – either the belief is unacceptably inevitable, or unacceptably chancy. Second, the presence of some flaws is well-correlated with the presence of other flaws. For instance, if the correct causal explanation of a belief doesn't presuppose or entail its truth, then it's likely that this belief will also be unacceptably inevitable. This is because if the moral facts don't figure into the explanation of that belief, then when we changed these facts, that belief would still remain the same. So if a belief suffers from the explanation-based flaw of being explainable without presupposing/entailing its truth, that belief will likely also suffer from the process-based flaw of being unacceptably inevitable. In contrast, this explanation-based flaw doesn't entail that the other process-based flaw of being unacceptably chancy – because even if a belief is explainable without reference to the moral facts, there might be some other cause (that isn't the moral facts) that still rigidly produced our moral beliefs anyway.

So much for the Epistemic Flaws. There are also further issues with each flaw: those who endorse process-based flaws need to tell us how to individuate processes;[12] those who like explanation-based flaws have to say what makes a good enough or admissible explanation;[13] those relying on disagreement-based flaws have to tell us what counts as epistemically significant disagreement,[14] and so on.

[12] Machery (2017, pp. 97–99) argues, however, that the generality problem for process reliabilism doesn't affect the use of process-based flaws in experimental philosophy, because we can circumvent it by looking for the most fine-grained process type whose reliability is invariant under further partitioning.

[13] Shafer Landau (2012, pp. 24–25) argues that the evolutionary explanations recruited by debunking arguments might be the best explanations we have and yet still fail to be evidentially compelling, because of our general lack of evidence about our evolutionary history. Also see Chapter 4.

[14] See Bogardus (2016, pp. 655–657) for some discussion of this; also see my Chapter 4, where I discuss disagreement-based flaws together with the process-based flaw of chanciness.

To conclude this section, consider an important distinction that cross-cuts the different flaws. Vavova (2015, pp. 105–107) distinguishes between an argument that demands good reason for thinking that our moral beliefs aren't mistaken, and one that gives good reason to think that our moral beliefs are mistaken. She draws an analogy to a skeptic who has done no tests and challenges us to show that our sense perception is accurate – contrasting that with an optometrist, who has done tests to show us that we are colourblind. Both the skeptic and the optometrist challenge our perceptual beliefs, but they do so in different ways – the skeptic *demands* good reason to think that our beliefs are not mistaken, whereas the optometrist *gives us* good reason to think that they are mistaken. Vavova argues that debunking arguments should be like the optometrist's challenge – using empirical evidence to give us good reason to think that the target moral beliefs are mistaken. Even if you disagree, I think she has proposed a useful distinction. For each flaw I've listed above, we distinguish between a challenge that demands good reasons for thinking that the target moral beliefs don't suffer from that flaw, from a challenge that gives us good reason to think that our beliefs do have that flaw. For instance, with regards to process-based flaws – some challenges demand good reason to think that our moral beliefs weren't produced by an unreliable process, while other challenges give us good reason to think that our moral beliefs were produced by an unreliable process.

1.1.4 The Empirical Evidence

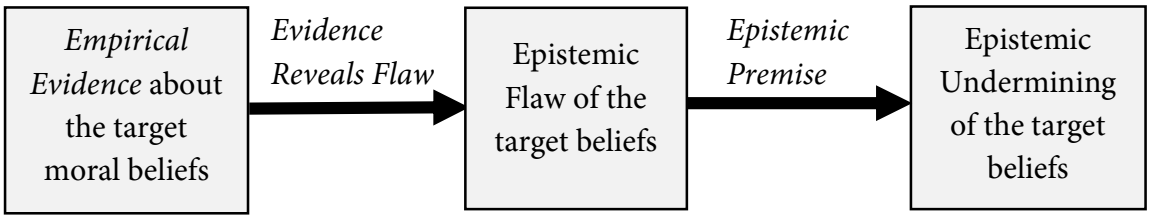
Finally, debunking arguments rely on some Empirical Evidence to show that the target beliefs suffer from the relevant Epistemic Flaw. Here are some examples of evidence invoked by debunking arguments:

- The ethnographic record, which documents the frequency of different kinds of moral norms, or the existence of moral concepts, in different societies and cultures
- Developmental evidence, for instance about children's ability to reason with deontic conditionals
- Cross-species comparisons about the presence of a trait (like seemingly moral behaviour) in other species

- Experimental evidence about how our moral intuitions or beliefs change under different conditions.

At this point, we don't have to go into detail about each piece of evidence.[15] Instead, just remember that some evidence concerns the nature of proximate causes – the immediate psychological causes of the belief or intuition – while others concern the ultimate causes, which stretch far back into history.

We've now seen the different ingredients of a debunking argument, and their different variations. These are summarised below:



- Ethnographic record
- Developmental evidence
- Cross-species comparisons
- Experimental evidence

- Process-based flaws
  - Actual unreliability
  - Counterfactual unreliability
- Explanation-based flaws
- Disagreement-based flaws

- Likely false
- Unjustified
- Have justification reduced
- Doesn't count as knowledge

1.2 Objections to Debunking Arguments

Now let's turn to potential objections to debunking arguments.

1.2.1 Contesting the Flaw/Third-Factor Accounts

Many prominent debunking arguments are *evolutionary* – they say that our moral beliefs are flawed because these beliefs were produced by problematic evolutionary causes like natural selection over many generations, or genetic drift.

[15] For a good review of the evolutionary evidence used in debunking, see Machery and Mallon (2010); for the experimental evidence, see Machery (2017, Chapter 2).

Some philosophers reply that the causal influence of evolution isn't problematic to begin with. They offer explanations for how evolutionary causes could have produced trustworthy moral beliefs. Enoch (2010, pp. 430–432), for instance, argues that evolutionary causes tend to produce beliefs that promote survival or reproductive success. And coupled with the moral fact that survival or reproductive success tends to be a morally good thing (even though its goodness might be outweighed, or it might only be instrumentally good for some other thing), Enoch thinks we can explain why our moral beliefs are somewhat correlated with the moral facts. By pointing to a third factor – the fact that survival is generally a morally good thing – he argues that we can explain how our evolutionarily-influenced moral beliefs could have lined up with the moral facts, without being causally or constitutively related to them. More generally, this kind of response runs as follows: we want to explain how some A-facts (that we hold some moral beliefs) are correlated with some B-facts (that these moral beliefs are true). We appeal to some third factor, the C-facts (that survival is generally morally good), which influences both the A- and B-facts, to explain their correlation. Importantly. The A- and B-facts do not have to be causally or constitutively related in any way – it's the third factor, the C-facts, that explain the correlation between them.

This kind of objection – known as a *third-factor account* – has been pursued in different ways. Copp (2008, pp. 198–202) argues that morality has the function of helping societies to meet their basic needs – a moral code ensures the continued existence of a society, and fosters peace and cooperation between its members and with neighbouring societies. He then argues that evolutionary causes – both biological and cultural – would have produced moral beliefs that were approximately true, since the beliefs produced would have helped societies to meet these basic needs. Wielenberg (2010, pp. 449–450) argues that for an organism to form moral beliefs about rights, it must have had sufficiently sophisticated cognitive faculties to do so. And if an organism had such cognitive faculties, it too would possess rights of its own. Thus we can explain why our moral beliefs about rights could be correlated with facts about the existence of such rights. Skarsaune (2011, pp. 233–235) argues that evolution made us value certain reproductively beneficial things – like nutrient foods, sex, and the welfare of one's

children – by making us take pleasure in these things. But it is also true, as a moral claim, that pleasure is generally good for a subject, and pain is generally bad. So evolutionary processes would have produced beliefs that were approximately true – since such processes would have made us value certain things, but at the same time imbued these very things with value by making us take pleasure in them. Brosnan (2011, pp. 43, 59–61) notes that evolution would instil in us the belief that cooperation is morally good, since this belief would induce us to help our group, which promotes reproductive fitness. He then argues that this helping behaviour is also morally good, because it promotes wellbeing. Thus we can explain why the evolutionarily-produced belief that cooperation is morally good could be likely true. More generally, Brosnan argues that we can easily explain the correlation between our moral beliefs and the moral facts – as long as the natural facts (about our beliefs) obtaining would *raise the probability* of the moral facts obtaining, we can say that evolutionary processes will tend to align our moral beliefs with the moral facts. Berker (2014, pp. 228–230) argues that we could explain the correlation between our moral beliefs and the moral facts with even more complicated structures – for instance, if a factor F casually promotes our having certain moral beliefs, another factor F* metaphysically grounds the moral facts (at least partially), and F and F* are causally related in an appropriate way, then we could get the required correlation between facts and beliefs.

In terms of my earlier framework, third-factor accounts contest the link between the *Empirical Evidence* about evolutionary causes[16] and the Epistemic Flaw – that is, they contest *Evidence Reveals Flaw*. These third-factor accounts propose a positive account of the

[16] Third-factor accounts have been given primarily in response to debunking arguments involving the ultimate, evolutionary causes of our moral beliefs. But it's interesting to consider whether such accounts could be offered in response to debunking arguments involving *proximate* psychological causes. Perhaps the replies to debunking arguments based on prospect theory's distinction between gains and losses might be instances of this. For example, Van Roojen (1999) and Dreisbach and Guevara (2019) argue roughly that subjects' responses to the Asian Disease Problem might be tracking morally relevant features of the situation, rather than being subject to irrational framing effects.

relationship between our evolutionarily-caused beliefs and the moral facts which entails that evolution would have produced approximately true moral beliefs that *aren't* flawed.[17]

One reply to third-factor accounts is that they rely on some substantive claim about the moral facts: Enoch relies on the claim that survival is generally morally good; Wielenberg on the claim that organisms with certain cognitive faculties have rights; Skarsaune on the claim that pleasure is generally good for a subject and pain generally bad; Brosnan on the claim that promoting wellbeing is generally morally good. A debunker might reply that assuming these claims in the context of replying to a debunking argument is just question-begging, since it assumes that we have trustworthy access to the moral facts to begin with. Street (2006, pp. 135–141), for instance, contends that such replies simply reassert the correlation between the moral facts and our moral beliefs, without explaining the correlation. She also argues that we can simply re-run a further debunking argument against our beliefs about these substantive assumptions – these beliefs will then turn out to be epistemically untrustworthy as well.[18] Copp (2018) concedes that third-factor accounts beg the question in the context of convincing a debunker – because in this context, the common ground includes the claim that all our moral beliefs are unjustified. He allows, however, that third-factor accounts don't beg the question in a different context, where we assume some moral claims, and are just trying to explain the potentially puzzling correlation between our moral beliefs and the moral facts. Berker (2014, p. 249), on the other hand, argues that Street's demands manifest a perfectly general sceptical worry – about how we can show that our cognitive faculties are reliable without relying on these faculties themselves to do so. He argues that this problem might have no satisfactory solution, that it is not special to our moral faculties, and that it does not make essential reference to the evolutionary evidence anyway.

This debate around third-factor accounts touches on an important issue: what we are entitled to assume when entertaining debunking arguments against our moral beliefs. I'll touch on

[17] Chapter 4 examines a different way of contesting *Evidence Reveals Flaw* that doesn't rely on such strong claims about what the moral facts are like.

[18] Street's target – mind-independent moral realism – is more restricted than I'm making it out to be, but the spirit of her replies can be extended to all third-factor accounts (Berker, 2014, n. 39).

this issue throughout the thesis, but particularly so in Chapter 3, though in relation to a different kind of debunking argument. For now, let's move on to a related reply to debunking arguments.

1.2.2 Global debunking arguments cannot provide good reason to doubt our moral beliefs

The debunker says that we are not entitled to assume a substantive moral view – like Enoch, Skarsaune, Wielenberg and Brosnan have – in answering a debunking argument, since doing so would be question-begging. Vavova (2014, pp. 89–93) argues, however, that this restriction might disable global debunking arguments themselves too.

Recall that global debunking arguments purport to undermine all our moral beliefs at once. Recall, too, that Vavova argues that debunking arguments should – like an optometrist, rather than an armchair skeptic – give us good reason to think that our moral beliefs are flawed (as opposed to merely demanding good reason to think that they aren't flawed). If we agree with her on this, then it seems like global debunking arguments might be question-begging too. Because if we are not entitled to assume a substantive moral view, then we have no idea what morality is like. It might be that the moral facts are exactly as we believe them to be, or it might be that our moral beliefs are completely inaccurate – we just can't tell, since we cannot assume anything about what the moral facts are like. In this case, then, we can't have good reasons for thinking that our moral beliefs are flawed either, so the global debunking argument doesn't work.

Vavova (2014, pp. 95–100) is more optimistic, however, about local debunking arguments – which target some, but not all, of our moral beliefs. By targeting only some moral beliefs, the local debunking argument allows for some other beliefs to provide independent ground from which to launch a debunking argument. She proposes an inverse law of debunking, which states that the debunking argument's impact on our moral beliefs is inversely proportional to

the degree to which it impacts crucial beliefs that provide the independent ground needed to evaluate the evidence.

1.2.3 Proving Too Much/Companions in Guilt

Another potential issue is that debunking arguments prove too much. The thought here is that if we brought arguments of the same form to our non-moral beliefs, then such beliefs would be similarly undermined. In particular, our non-moral beliefs might suffer from the same Epistemic Flaws alleged of our moral beliefs. Some philosophers invoke these 'companions in guilt' in the hopes of showing that debunking arguments either hold our moral beliefs to an overly demanding standard, or that such arguments don't pose a special problem to our moral beliefs.[19]

Some common companions in guilt are listed below, along with the Epistemic Flaw they allegedly share with our moral beliefs.

Companion in Guilt	**Alleged Epistemic Flaw**
Empirically undetermined beliefs	Counterfactually unreliable (inevitable) (White, 2010, p. 591)
Beliefs about the future	Available debunking explanation (Nichols, 2014, p. 728; White, 2010, pp. 582–583)
Beliefs about necessary facts	Counterfactually unreliable (inevitable); Available debunking explanation (White, 2010, pp. 582–583)
Beliefs about complicated topics like linear algebra, theoretical physics, evolutionary biology, and epistemological reflections[20] on	No vindicating evolutionary explanation (White, 2010, pp. 591–592)

[19] I believe these are actually two different ways of responding, since the first seems to rule out skepticism about all our moral beliefs, whereas the second doesn't.

[20] Srinivasan (2015, pp. 328–342) outlines how the epistemic premises of different debunking arguments are subject to the very same epistemic flaws alleged by the argument (though she doesn't think this strategy is sufficient for defending the target judgments, see the next section for why). Street's (2009) evolutionary debunking argument – which is meant to support constructivism over realism – also affects beliefs about epistemic reasons.

evolutionary biology (as opposed to basic logic, arithmetic, induction)	
Perceptual beliefs	Counterfactually unreliable (chancy) (Shafer-Landau, 2012, n. 21)
All our other beliefs	Counterfactually unreliable (inevitable) (Shafer-Landau, 2012, p. 18; White, 2010, p. 581)

The companions-in-guilt move questions *Epistemic Premise* – the link between the Epistemic Flaw and the Epistemic Undermining. If we think that any of these companions-in-guilt beliefs have positive epistemic status, then we cannot think that the corresponding Epistemic Flaw is sufficient for undermining a belief. While a popular move, it still seems open for the debunker to just admit that these companions in guilt are similarly undermined too, so it's unclear if companions in guilt will help the opponent of debunking much.

It's also hard to assess the plausibility of the companions-in-guilt move because of its lack of specificity. The move compares entire swathes of beliefs – for instance, *all* our moral beliefs with *all* our perceptual beliefs – without considering more details about each token belief. Without these details, it's difficult to assess and compare the epistemic status of beliefs in different domains, and to determine whether we should accept or reject some specific epistemic standard. Because of this, I'll set this move aside for the rest of this thesis. However, in Chapter 2, I focus on a structurally similar move: other philosophers object to Epistemic Premise using counterexamples from epistemology (Bogardus, 2016, pp. 644, 649–650; Sinclair, 2018, pp. 106–107; White, 2010, pp. 580–581).

1.2.4 Self-Defeat

A debunking argument might even prove so much that it undermines beliefs in *its own premises*. For instance, the *Epistemic Premise* of a debunking argument is a normative claim about what we should believe, and beliefs about this premise might be similarly undermined by the evolutionary findings (Srinivasan, 2015, pp. 328–342). So it might be that a debunking argument, if successful, undermines its own premises and hence cannot work.

Srinivasan (2015, p. 346) points out, however, that the mere fact that the debunking argument is self-defeating is insufficient for defending the target moral beliefs. This is because we might still be committed to the premises of the debunking argument anyway. In this case, the skeptic is using our own commitments against us – and it's our own belief system, rather than the debunking argument itself, that is self-defeating. Put differently, even if the debunking argument is self-defeating, it still works to highlight a tension within our own commitments. We can think of it as a *reductio ad absurdum* of anyone who wants to endorse both the premises of a debunking argument, and the moral beliefs targeted by it. I believe Srinivasan is right. In Chapter 2, I develop her ideas by distinguishing between (what I call) personal and adversarial views of debunking, and by considering and rejecting further ways in which an opponent of debunking might escape this *reductio* of their belief system.

1.2.5 The Regress Objection

A related but distinct objection contends that the debunking argument commits us to a problematic regress. Roughly, the idea is this: in order to show that a moral belief is epistemically flawed, we need to rely on some other moral beliefs as assumptions. But how do we know that these relied-upon beliefs aren't also flawed? We need to show that these relied-upon beliefs have positive epistemic status. But showing this requires a further set of moral beliefs, whose epistemic status needs to be confirmed by yet another set of moral beliefs, and so on. Regina Rini (2016) and others argue that the supporter of a debunking argument is committed to an insurmountable regress, and that this disables the debunking conclusion. In Chapter 3, I explore this objection in detail, and try to defuse it – so I'll leave detailed treatment of the regress till then.

1.3 Further Issues

We've looked at debunking arguments and some potential objections. Now let's explore two further issues.

1.3.1 Metaphysical Conclusions and Assumptions in Debunking

Some debunkers go further to draw metaphysical conclusions from debunking arguments. Most famously, the conclusion of Sharon Street's (2006) debunking argument is that the moral facts are mind-dependent – that is, such facts constitutively depend on a person's evaluative attitudes. A simplified reconstruction of her argument runs as follows:

> (P1) If the moral facts are mind-independent and our moral beliefs are thoroughly influenced by evolutionary processes, then all our moral beliefs are unjustified.
>
> (P2) At least some of our moral beliefs are justified.
>
> (P3) Our moral beliefs are thoroughly influenced by evolutionary processes.
>
> (C1) Therefore, the moral facts are not mind-independent (that is, they are mind-dependent). (from P1, P2, P3)

Much of what I've been calling debunking so far could support a conditional claim like P1. But Street goes further to claim that at least some of our moral beliefs are justified, despite the evolutionary influence. Thus, she thinks, we should adopt a different view of the metaphysical grounds of the moral facts – we should think that such facts are mind-dependent. Tropman (2014), Berker (2014), FitzPatrick (2015, n. 10), and Joyce (2016, pp. 127–128) argue, however, that even if we accept that the moral facts are mind-dependent, this might not restore justification to our moral beliefs. Joyce gives the example of someone rolling a die to decide what to believe about the currency exchange rates in a foreign country – he argues that the resulting belief is rendered unjustified by its origins, even if this belief is about a mind-

dependent fact. I agree with these critics – the debunking argument doesn't give us reasons to think the moral facts are mind-dependent rather than mind-independent. So for this thesis, I will not explicitly discuss Street's metaphysical conclusions. Nonetheless, everything I say later about evolutionary debunking – particularly in Chapter 4 – will still pertain to Street, since my arguments might still weaken conditional claims like P1. And to that extent, anyone sympathetic to Street would still find something of value in this thesis.

Instead, my chief target when it comes to evolutionary debunking will be Joyce (2006, 2016). He argues that the discovery of evolutionary influences would impact the epistemic status of our moral beliefs – and that this conclusion holds across many different views about what the metaphysical grounds of the moral facts are.

Even if Joyce's conclusions are epistemic, the force of these conclusions might in fact stem from the strong metaphysical assumptions of his argument (Das, 2016, pp. 422–424) – for instance, about the impossibility of reducing the moral to the non-moral (Joyce, 2006, pp. 188–199). I agree that some epistemic conclusions might derive their force from strong metaphysical assumptions – but I also believe that other epistemic conclusions might not, and instead could be obtained while assuming something weaker about the moral metaphysics. Throughout this thesis, I will focus first on the epistemic conclusions in debunking, but then also highlight when I think metaphysical assumptions are required to reach them.

1.3.2 Empirical Evidence about Proximate vs Ultimate Causes

As mentioned earlier, debunking arguments can rely on two broad kinds of evidence to undermine our moral beliefs: evidence about the *proximate* causes of moral belief – that is, the immediate psychological mechanisms that produced our beliefs or intuitions; or evidence about their *ultimate* causes – that is, historical causes such as natural selection operating over many generations. We can see the ultimate causes – like evolution – as producing proximate causes, which then produce our intuitions and beliefs. This can be represented in the (oversimplified) structure below:

Ultimate causes (like evolution) → Proximate causes (immediate psychological mechanisms) → Intuitions/Beliefs

O'Neill (2015, pp. 1072–1074) argues that once we recognise this causal structure, we can see how information about proximate causes is more valuable for inferring epistemic status than information about ultimate causes.[21] This is because ultimate causes like evolution only affect our moral beliefs *through* the proximate causes – evolutionary processes give rise to developmental processes, which then produce the proximate psychological mechanisms that produce moral intuitions and beliefs. If we had credible information about whether proximate causes produced trustworthy moral beliefs, then further information about whether evolution produces truth-tracking psychological mechanisms wouldn't tell us anything more about the epistemic status of these beliefs. On the other hand, if we only had information about the ultimate evolutionary processes, learning new information about the proximate psychological causes would tell us something more about the trustworthiness of our moral beliefs. I believe this can be understood as a point about the screening-off of one kind of cause by another – I outline this as part of my Bayesian analysis in Chapter 6.

Mogensen (2015) argues that the debunkers mistakenly move from claiming that the moral facts are not part of the ultimate causes of moral belief to the claim that such facts are not part of the proximate causes either. FitzPatrick (2016, p. 436) responds that the debunkers think "it's morally blind causes all the way down" – both the proximate and ultimate causes of our moral belief don't track the moral facts in any way. He allows, however, that the debunkers might have overlooked the role of alternative proximate causes – reflection and reasoning – which could have produced trustworthy moral beliefs.

[21] O'Neill (2015, n. 1) defines proximate and ultimate causes in a different way than I do – but I believe these don't matter for my interpretation of her argument.

1.4 Conclusion

In this chapter, I've examined debunking arguments, which comprise of three main premises – *Empirical Evidence*, *Evidence Reveals Flaw*, and *Epistemic Premise*. I then surveyed possible variations in the different parts of a debunking argument – to do with the varieties of epistemic undermining, the types of epistemic flaws, and the different kinds of empirical evidence that could be invoked. I also looked at possible objections to such arguments: third-factor accounts which contest the epistemic flaw to begin with; the inability of a global debunking argument to provide good reasons to think that the target beliefs are flawed; companions in guilt which seem to have positive epistemic status despite having the same epistemic flaws alleged of our moral beliefs; the potentially self-defeating nature of a debunking argument; and the regress objection. Finally, I examined some further issues – concerning the metaphysical conclusions and assumptions in debunking, and concerning the distinction between evidence about proximate and ultimate causes. With this background in place, we can start exploring substantive issues – in the next chapter, I defend the in-principle viability of debunking arguments against counterexamples to their *Epistemic Premise*.

2 The Epistemic Premise of Debunking Arguments

Debunking arguments pose an epistemic challenge to the target moral beliefs. A crucial part of this challenge is their *Epistemic Premise*, which embodies the standards used to evaluate the target beliefs:

(*Epistemic Premise*) If a belief has flaw F, then it is epistemically undermined.

Opponents of debunking have argued against this premise, contending that various versions of it are subject to counterexamples: beliefs which suffer from some epistemic flaw, but which aren't epistemically undermined nonetheless. If these opponents are right, then debunking arguments cannot work in principle, because such arguments are based on erroneous epistemic standards. In this chapter, I respond to these opponents, and argue instead that debunking arguments are in principle viable. My arguments have two main strands. The first (sections 2.1-2.3) examines prominent counterexamples – I argue that these either pertain to epistemic standards that are inappropriate for moral methodology, or stem from misunderstandings about what counts as an epistemic flaw. Through my arguments, I clarify the exact versions of *Epistemic Premise* that should figure in debunking arguments.

The second strand (sections 2.4-2.5) argues more generally that we cannot just deny *Epistemic Premise* – because this makes it difficult to explain why some evidently problematic beliefs are indeed epistemically problematic. To properly execute the strategy of denying *Epistemic Premise*, an opponent of debunking must propose a replacement premise that disqualifies the clearly problematic cases, without also posing a skeptical threat to the moral beliefs they wish to defend. I consider three replacement versions of *Epistemic Premise* which aim to limit the potential scope of the debunking argument in order to defend the target beliefs – and argue

that all are implausible. In the end, I conclude that if we are to resist a debunking argument, it should not be through denying its *Epistemic Premise*.

2.1 Counterexamples to Epistemic Premise

Our beliefs – moral or otherwise – can be flawed in different ways. Debunking arguments pick on an alleged flaw of the target moral beliefs, and use this flaw to undermine these beliefs. In this thesis, I'll focus on the prominent process-based flaws of chanciness and inevitability.[1] And, in this section, I outline debunking arguments which rely on these flaws, paying special attention to debates over their *Epistemic Premise*.

2.1.1 Chanciness-Based Debunking

Start first with the epistemic flaw of chanciness. Recall that a belief is chancy if it could easily have been different, given different belief-forming circumstances – even when the facts (that the belief is about) remain the same. Many debunking arguments allege that our moral beliefs are flawed in just this way. For instance, some evolutionary debunking arguments contend that we could easily have evolved differently, and had correspondingly different moral beliefs, or no moral beliefs at all (Darwin, 1871, p. 102; Joyce, 2006, p. 181; Ruse & Wilson, 1985, p. 52).[2] The idea here is that our moral beliefs have been greatly influenced by our specific evolutionary history, and are highly contingent on how this history has played out. Had this history gone differently (and it could easily have), we would have ended up with very different moral beliefs, even when the moral facts remain the same. Debunking arguments based on proximate causes often also invoke chanciness, but at a different temporal scale: such arguments might allege, for instance, that we could easily have encountered a thought experiment presented in one order rather than another, and formed a different belief about it

[1] I'm using the terms 'chanciness' and 'inevitability' instead of the more familiar terms of being unsafe, insensitive, or non-adherent. This is because these more familiar terms are associated with the study of knowledge and justification, whereas I'm more concerned with studying the probability that our moral beliefs are true (see the later sections for how these come apart).

[2] Bogardus (2016, pp. 645–647) also sees some debunking arguments as relying on chanciness.

– even though the moral facts stay the same across different orders of presentation. These arguments all rely on the same core idea: that the target moral belief is unacceptably chancy or contingent, since it could easily have changed across different belief-forming circumstances, even when the moral facts remain the same.

The debunkers use non-moral analogies to argue that chancy beliefs are epistemically undermined. For instance, Richard Joyce (2006, 2016) likens our evolutionarily-produced moral beliefs to beliefs that were produced by a Napoleon belief pill.[3] Suppose you believe that Napoleon exists. You then learn that you were the subject of an experiment, and were unknowingly fed one of two pills chosen at random: one pill causes you to believe that Napoleon exists, and another pill causes you to believe that he doesn't. You could easily have been fed the pill that causes you to believe that Napoleon doesn't exist instead – in which case, you would have very different beliefs, even when the facts about Napoleon's existence remain the same. Learning about this chanciness seems to undermine your actual belief that Napoleon exists – you should think that this belief is less probable than before to be true.[4] We can put this in terms of the following *Epistemic Premise*: "If a belief is chancy, then we should think that this belief is less probable than before to be true." I'll use the following shorthand, of the form 'flaw → undermining', to express this version of *Epistemic Premise* (we'll consider several more versions throughout this chapter) –

Chancy → Lower probability of truth

[3] Joyce (2006, p. 181) gestures at problematic chanciness in our evolutionary history when he says that "Were it not for a certain social ancestry affecting our biology, the argument goes, we wouldn't have concepts like obligation, virtue, property, desert, and fairness at all". There are a few different ways of filling out the belief pill thought experiment, which lead to examples of either chancy or inevitable beliefs. In what follows, I will explicitly set out two different versions of the thought experiment, one for chanciness and one for inevitability.

[4] Joyce has argued for different kinds of epistemic undermining in his work – some of which concern probability of truth; others concern justification. For instance, he has argued that recognising the contingent evolutionary origins of our moral beliefs means that we should be agnostic about such beliefs (Joyce, 2006, p. 181), or should reduce confidence in them (Joyce, 2016, p. 125) – these pertain more to probability of truth. On the other hand, he has also argued that this contingency means that our moral beliefs are rendered unjustified (Joyce, 2006, p. 180). I believe that these are substantively different positions that should be kept apart – I explain why in the following sections.

In response, opponents of this debunking argument use other cases to argue that a belief's being chancy doesn't disqualify it from *being justified*, or from *counting as knowledge* (Bogardus, 2016, pp. 649–650; Sinclair, 2018, pp. 106–107). These opponents are also concerned with chanciness, but they are thinking about its leading to a different kind of epistemic undermining – of undermining justification or knowledge. For instance, Bogardus (2016, pp. 649–650) presents a case that seems to show how a belief could be chancy[5] and yet still count as justified or knowledge:

> **Atomic Clock:** Smith forms his beliefs about the time using his atomic clock, which he knows to be the most accurate clock in the world. However, this clock is very sensitive to nearby radiation – if radiation is present, the clock will malfunction and display an incorrect time. As it happens, there is a radioactive isotope nearby – it has a chance of decaying, emitting radiation and causing Smith's clock to malfunction. Smith glances at the clock and forms a belief that the time is 8:22. Luckily for him, the isotope did not decay and his belief is true. However, the isotope could easily have decayed and caused a malfunction.

In this case, we might intuitively judge that Smith's belief is justified, and that it counts as knowledge – even though this belief is chancy (Bogardus, 2016, pp. 649–650). The Atomic Clock case thus contests the *Epistemic Premise* that

> Chancy → Not justified/knowledge

[5] Bogardus understands the relevant epistemic requirement (in this case and in chanciness-based debunking arguments generally) to be *safety*. On one common construal, S's belief that p is safe iff if S were to believe that p, p would be true (Sosa, 1999, p. 146). This looks at nearby worlds where S believes that p, and then checks whether p is true in those worlds. I think we shouldn't understand chanciness-based debunking arguments in terms of safety, however – because such arguments rather concern nearby worlds *where the moral facts are the same* (that is, where p is true), and where we then check whether we still believe that p in those worlds. So the requirement is rather that: if p were true, S would believe that p. This is Nozick's (1981, p. 176) adherence condition, which is also cited by O'Neill (2015) in the context of debunking. Also see Roush (2007, pp. 121–123), who argues that the safety requirement gets the direction of fit wrong – it requires the world to fit the belief – and is hence unsuitable as an epistemic requirement.

On one hand, the debunkers give a case where learning about a belief's chanciness should lead to an impact on that belief's probability of truth. On the other hand, their opponents cite a different case which seems to show that a chancy belief can still be justified, or still count as knowledge. What's going on here? Before I try to resolve this, I'll outline how similar moves have played out with the epistemic flaw of inevitability.

2.1.2 Inevitability-Based Debunking

Recall that a belief is inevitable if had the facts (that the belief is about) been different, that belief would still remain the same. The idea here is that the belief is too insensitive to the facts, or that it is too rigid. There are also debunking arguments based on the flaw of inevitability. Some argue that our evolutionarily-produced moral beliefs are inevitable, because evolution has forcefully inculcated these beliefs in us – such that had the moral facts been different, we would still have had the same moral beliefs, because these beliefs would still promote reproductive fitness (Morton, 2016, pp. 242–243; Ruse & Wilson, 1986).[6] Learning about this inevitability, the debunkers think, should also undermine our moral beliefs.

Again, the debunkers can draw a non-moral analogy to support their argument. Consider a variation of the belief pill case: you believe that Napoleon exists, but now you learn for sure that you were slipped a believe-that-Napoleon-exists pill in the past – this pill causes you to form Napoleon beliefs. So had Napoleon *not* existed, the pill would still have ensured your belief that he did exist. Learning about this inevitability should also lead to an impact on the probability that your belief is true – perhaps you should also think that your belief is less probable than before to be true. This case might thus support the following *Epistemic Premise*:

Inevitable → Lower probability of truth

[6] Kahane (2011, p. 106), Clarke-Doane (2012), and Bogardus (2016, pp. 638–640) also view some debunking arguments as relying on the flaw of inevitability.

However, the opponents of debunking also gesture at counterexamples showing how an inevitable belief can still be justified or still count as knowledge (Bogardus, 2016, p. 644; Sinclair, 2018, pp. 106–107; White, 2010, pp. 580–581). One such counterexample might be Sosa's (2000, p. 13) garbage chute case.[7] It's not important to go into the specifics of that case here, but the idea should be clear – these opponents contend that there are counterexamples to the premise that:

Inevitable → Not justified/knowledge

The debunkers allege that the target moral beliefs suffer from some flaw F – being chancy, or being inevitable – and conclude that these beliefs should have a lower probability of truth. In response, their opponents offer counterexamples to a different version of *Epistemic Premise* – arguing that beliefs with flaw F could still be justified and still count as knowledge. The debunkers and their opponents are talking past each other – they are concerned with different kinds of epistemic undermining. To settle the debate, we'll need to decide which kind of undermining we should be concerned with in debunking, and hence whose cases are more appropriate. In what follows, I'll argue that if our aim is to improve moral theorising, we should be concerned with the probability that our moral beliefs are true, rather than with their status as justified or as knowledge. Consequently, the debunkers' cases, and their corresponding versions of *Epistemic Premise*, are more appropriate – and their opponents' cases should pose no trouble for the debunking argument.

[7] The case is as follows: I throw a trash bag down the garbage chute of my high-rise condo. I believe that the trash bag landed in the basement – and the trash bag did indeed land there, so my belief is true. However, if the trash bag hadn't landed in the basement, it would have been snagged somewhere in the chute, and I would still have believed that it landed in the basement. This belief is inevitable, but it might still seem to be justified, or count as knowledge.

2.2 Which Kind of Epistemic Undermining?

Notice first that the different possible kinds of undermining created by a debunking argument – concerning justification, knowledge, and probability of truth – correspond to three different projects we could undertake in epistemology. These ask distinct questions about the epistemic status of our beliefs:

> Justification Project: Which of our beliefs are justified? What are the conditions for a belief to be justified?
>
> Knowledge Project: Which of our beliefs count as knowledge? What are the conditions for a belief to count as knowledge?
>
> Probability of Truth Project: How likely is it that our beliefs are true? What are the conditions that render a belief more or less likely to be true?

Much of the debunking literature (and of moral epistemology more generally) has focused on the Justification and Knowledge Projects.[8] But I believe that if we were interested in moral methodology, we should focus on the Probability of Truth Project instead. Here's why: when doing moral theorising – into, say, whether there are any moral facts at all, or whether eating factory-farmed meat is morally forbidden – I'm not interested in obtaining a justified belief about these matters for its own sake; neither am I aiming for moral knowledge in the first instance. Instead, I'm aiming to get at the *truth* about these matters – for instance, about whether eating factory-farmed meat is *in fact* morally forbidden or not. So, when evaluating my moral beliefs for use in moral theorising, I should evaluate these beliefs based on their probability of truth, rather than on their being justified or counting as knowledge. That is, if our concern is with improving moral theorising, we should be engaged in the Probability of Truth Project, and not in the Justification or Knowledge Projects. More specifically, if we

[8] For instance, see Joyce (2006), Wielenberg (2010), Kahane (2011), Brosnan (2011), Clarke-Doane (2012), Fraser (2014), Nichols (2014), FitzPatrick (2015), Bogardus (2016), Mogensen (2016a), Morton (2016), Sinclair (2018).

were interested in debunking arguments for their potential to improve moral theory, we should be concerned with their impact on probability of truth, rather than on justification or knowledge.[9]

This conclusion wouldn't seem significant if you thought that the results of the Justification and Knowledge Projects are good enough proxies for the results of the Probability of Truth Project – such that once we can determine whether and when a moral belief is justified or knowledge, we'll also be able to determine whether and when that belief is probably true. I think, however, that there are important ways in which the Knowledge and Justification Projects come apart from the Probability of Truth Project. And when they do, the results of the Justification and Knowledge Projects will be uninformative or even misleading for moral methodology. I'll first give theoretical arguments for why. I then illustrate how this implies that cases like Atomic Clock are inappropriate for informing the *Epistemic Premise* of debunking arguments.

First think about the Justification Project, which is often conducted by explicating our ordinary concept of justification. Philosophers elicit intuitions about when token beliefs count as justified, and try to discern the conditions under which a belief counts as justified, according to our ordinary concept. The problem here is that these conditions and intuitions are at best orthogonal to the purpose of obtaining true moral beliefs and avoiding false ones – our ordinary concepts might have been shaped such that the conditions for justification only partially overlap with the conditions for likely truth (Shogenji, 2012, p. 61).[10] Alston (1985, pp. 67–68) notices, for instance, that our ordinary concept of epistemic justification involves satisfying our epistemic duties or obligations – such that a subject's belief is justified iff they are not violating any epistemic duties or obligations in holding that belief. But a subject could satisfy these duties and obligations while still holding a belief that is likely false. Alston gives

[9] This is not to say that the Justification and Knowledge Projects cannot be the proper focus for other kinds of moral epistemology. These projects might be relevant for those interested in how much of the folk's moral beliefs are held in an epistemically responsible way, for instance. Also, I'm happy to still concede that knowing that p is more epistemically valuable than merely having a true belief that p.

[10] The Knowledge Project might also be vulnerable to this problem, since it involves explicating our ordinary concept of knowledge. But I think it faces a more serious worry, which I focus on below.

the example of someone born in a culturally isolated tribe – accepting its unfounded traditions as authoritative, and never encountering any evidence that casts doubt on them. This person might have done everything that could be reasonably expected of them when forming their beliefs, and thus have discharged their epistemic duties. So their beliefs might be justified, according to our ordinary concept. But their beliefs are still likely false, given all the information available to us. If there are moral beliefs like this, we would want our moral methodology to disqualify them from our moral theorising – indeed, such beliefs are precisely of the kind that the debunkers are worried about. But if our moral methodology was using the ordinary concept of justification, it wouldn't be able to disqualify these beliefs. So the ordinary concept of justification, as studied by the Justification Project, is too weak for use in moral methodology.[11]

Now think about how this applies to Atomic Clock, where we had the intuition that Smith's chancy belief is justified (quite apart from its counting as knowledge). I believe this intuition is at least partly driven by our ordinary concept of justification. This is especially so if we assume that Smith doesn't have information about the isotope in the vicinity – in which case Smith might be justified because he did his best in using a clock that seems to be the most accurate in the world. But Smith's belief being justified is compatible with our having information that his belief has a significant probability of being false – information about the isotope and the chancy circumstances it creates, for instance. This illustrates how thinking in terms of the ordinary concept of justification might be too weak for moral methodology – since a belief can be justified according to this concept, even though we have information that it's likely false. (Of course, in Atomic Clock, this information is overridden by the case's additional stipulation that the isotope didn't in fact decay, and that Smith's belief is in fact

[11] There is an alternative, 'non-deontological' conception of justification which might work better for moral methodology – this says that a subject is justified in holding a belief iff they believe in a way that makes it sufficiently likely that their belief is true (Steup, 2018, sec. 3.1). However, it seems clear that our epistemic intuitions are at least partly influenced by the deontological conception of justification – and to that extent, using intuitions about justification would be misleading.

true. I believe this stipulation is illegitimate – see the following discussion on the Knowledge Project for why.)

Now consider the Knowledge Project, which investigates the conditions for a belief to count as knowledge. Knowledge requires true belief, so inquiring into the conditions for knowledge might also yield conditions for true belief. The problem, however, is that the Knowledge Project is usually conducted by asking: what *more* do we need to add to a true belief, in order for this belief to count as knowledge?[12] When viewed this way, the results of the Knowledge Project will be mostly uninformative for a moral methodology concerned with obtaining true moral beliefs and avoiding false ones. The inquiry of the Knowledge Project usually *assumes* a true belief to start with – so the conditions it yields, and its verdicts about the relative significance of different conditions, will not be directly useful for obtaining true moral beliefs. An analogy will help illustrate this. Suppose I want to find out how to swim a freestyle lap in under a minute. You offer me a guide aimed at expert swimmers who already know how to swim a lap in under a minute – this guide teaches them how to qualify for Olympic swimming. This guide will not be directly useful to me, for two reasons: first, it's pitched at a completely different goal to the one I'm interested in – qualifying for the Olympics, as opposed to swimming a lap in under a minute. Secondly, its guidance is explicitly offered to those who have *already achieved* my goal of swimming a lap in under a minute, so it will likely exclude many of the things I need to do to achieve that goal in the first place. Now think back to moral methodology: looking to the Knowledge Project to guide our moral inquiry is like perusing this Olympic qualification guide for directions on how to swim a lap in under a minute. First, the Knowledge Project is concerned with a very different kind of epistemic status – knowledge, rather than true belief. Second, the Knowledge Project often assumes a true belief to start with, so its results wouldn't offer direct guidance on how to obtain true

[12] Ichikawa and Steup (2016) note that "Most epistemologists have found it overwhelmingly plausible that what is false cannot be known". Consequently, most analyses of knowledge – like the ones surveyed in Ichikawa and Steup (2016) – and critiques of such analyses – like Gettier (1963) – start off just specifying that the beliefs in question are true.

moral beliefs – in fact, it might even wrongly exclude many conditions that are crucial for promoting true beliefs in the first place.

These points diminish the significance of the intuition, in Atomic Clock, that Smith's chancy belief counts as knowledge. If we took this intuition on board, we would conclude that a belief's chanciness doesn't influence whether it counts as knowledge. But this conclusion doesn't imply anything about the importance of chanciness in determining the probability that a belief is true – because true belief is a very different epistemic goal from knowledge. Moreover, Smith's belief in Atomic Clock was specified to be true to begin with – such that we're inquiring what other conditions his true belief must satisfy in order to count as knowledge. Such inquiry might exclude conditions that bear on the probability of truth of the belief – conditions involving an absence of chanciness, for instance. These issues with Atomic Clock are most clearly expressed in terms of probabilities. Consider some belief that p. Let *knowledge* be the proposition that "This belief counts as knowledge", let *true* be the proposition that "This belief is true", and let *chancy* be the proposition that "This belief is chancy". I believe that Atomic Clock, because it specifies a true belief to start with, at best only shows that Pr(*knowledge*|*true*&*chancy*) = Pr(*knowledge*|*true*&~*chancy*) – that chanciness doesn't make a difference to whether a true belief counts as knowledge or not. Even if this is correct, it might still be that Pr(*true*|*chancy*) < Pr(*true*|~*chancy*) – chanciness could still reduce a belief's probability of truth – and it is this probability of truth that we're concerned with.

These same points apply to intuitions showing how an *inevitable* belief could still be justified or still count as knowledge. First, our intuitions about justification might be unduly swayed by the ordinary concept of justification, such that a subject's belief can be justified even if we have information that it's significantly likely to be false (information about its inevitability, for instance). Second, our intuitions about knowledge concern a completely different epistemic goal which shouldn't be the focus of moral methodology. And, given how inquiry about knowledge often assumes a true belief to start with, we might wrongly conclude that

avoiding inevitability is not useful for obtaining true moral beliefs, when in fact, it's only not useful for obtaining knowledge.

In sum, when analysing debunking arguments, we shouldn't think in terms of knowledge and justification, and we shouldn't be using intuitions about cases like Atomic Clock.[13] Such cases are at best uninformative for moral methodology, and at worst, they mislead us about the epistemic significance of flaws like chanciness and inevitability. We should instead think in terms of the Probability of Truth Project, and ask: how likely is it that our moral beliefs are true? What are the conditions or indicators that render our moral beliefs more or less likely to be true?[14] Chanciness and inevitability seem to be such indicators, as seen from the debunker's cases – these cases are the appropriate ones to use if we are concerned with moral methodology. In the next section, I'll further clarify what kinds of chanciness and inevitability could count as epistemic flaws.

Before moving to that, however, one important detail of the Probability of Truth Project needs to be specified: what kind of information should be used to estimate the probability that the target moral belief is true? I have so far spoken about 'our' information, but this is vague: it might refer only to information accessible by the individual agent holding the target moral belief, mirroring a similar requirement imposed by access internalists about justification. Or it could refer to full information about what the facts are like – this mirrors an externalist view of justification. If our concern is with improving moral theorising, then neither version is quite right. We shouldn't just use information accessible to the individual

[13] Alternatively, if we modified cases like Atomic Clock to fit the purposes of moral methodology, we will see that flaws like chanciness do lead to an impact on probability of truth. Consider a variant of Atomic Clock where we don't have information about whether the isotope decayed, or about whether Smith's belief is true. We learn that there is an isotope in the vicinity, and that there is a significant chance that it could have decayed. In this case, I think this information should have a significant impact on the probability that Smith's belief is true.

[14] This perspective is closely aligned with Bayesian epistemology – the branch of epistemology that studies rational constraints on our degrees of belief (or credences), which model our estimates of how likely it is that a proposition is true. In Chapter 6, I present a Bayesian analysis of the debunking debate. For now, we can just think in terms of epistemic premises that undermine the probability that the target moral beliefs are true. Also see Alston's (2006) epistemic desiderata framework, which recognises various epistemically relevant properties that contribute to the cognitive goal of maximising true beliefs and minimising false ones about matters of importance, and Christensen's (2007, p. 196) 'first-person perspective', which is the perspective of an agent using their own beliefs to improve their epistemic position.

agent, since there is often more information available to experts that could improve our probability estimates – and our moral methodology should draw on this information where possible. On the other hand, we cannot assess the probability that the target moral belief is true given full information, because such information just isn't available to us – we don't have access to a moral Oracle. Instead, I propose that we assess the probability of truth of the target moral belief, using all information accessible to human society at present. This draws on information that's potentially available to any expert in society – this is appropriately external from any specific individual, allowing moral methodology to criticise epistemically problematic beliefs. But the information used is also internal to human society as a whole, so we aren't under the unrealistic requirement to have full information about the truth of our moral beliefs.[15]

2.3 What Counts as an Epistemic Flaw?

In the previous section, I argued that if we were concerned with moral methodology, we should be concerned with the probability that the target moral beliefs are true, given all the information available to human society. But we don't have direct information about the truth of our moral beliefs, so we must instead look for indirect indicators to estimate their probability of truth. The two epistemic flaws we've seen – chanciness and inevitability – are such indicators, because the debunkers' cases support the following versions of *Epistemic Premise*:

Chancy → Lower probability of truth

Inevitable → Lower probability of truth

[15] This line of argument draws directly from Shogenji's (2012) arguments about internalism and externalism in general epistemology, although he does not speak specifically about estimating the probability of truth of our beliefs.

Aside from using non-moral analogies (like the belief pill cases) to support these premises, we can also give a theoretical argument for them. When we learn that a target belief is chancy or inevitable, what we learn is that the process responsible for that belief is disposed to produce false outputs in some relevant scenarios. So we should at least think that it's more probable than before that the target belief is among one of the false outputs. (Depending on the number of false outputs revealed, and the similarity between the relevant scenarios and our actual belief-forming circumstance, we might even think it's highly or sufficiently probable that our belief is a false output.[16]) So chanciness and inevitability should have some impact on the probability that the target moral beliefs are true.[17]

However, the opponents of debunking have given further counterexamples which purport to show that chanciness and inevitability aren't epistemic flaws at all. These counterexamples, if successful, will also threaten the in-principle viability of debunking arguments. I now explain and tackle such counterexamples – I argue that they only show that some kinds of chanciness and inevitability don't count as epistemic flaws, while leaving other kinds of chanciness and inevitability unscathed.

First consider White's counterexample, which purports to show that inevitability isn't an epistemic flaw:

> **Defendant's Guilt:** We believe that a defendant is guilty, and have overwhelming evidence for this belief – but the evidence nonetheless does not entail that belief. It is logically possible that we have all this overwhelming evidence, yet the defendant is not in fact guilty and our belief is false (White, 2010, pp. 580–581).

In Defendant's Guilt, it's logically possible that the facts of the matter are different, yet our belief about the defendant remains the same. Thus the belief seems, in some sense, to be

[16] For a more detailed discussion of how this assessment might proceed, see Alston (2006, pp. 109–112, 2006, Chapter 6).

[17] Notice also that when we don't have good information about whether our moral beliefs are true or not, the distinction between counterfactual and actual reliability collapses.

inevitable. But that belief isn't epistemically undermined as a result – for instance, upon learning about this inevitability in Defendant's Guilt, it doesn't seem like we should lower the probability that our belief about the defendant is true.[18] I think this is correct, but it merely shows that inevitability of some kind is epistemically unproblematic. In particular, in Defendant's Guilt, it is only *logically possible* that our belief remains the same, even when the facts of the matter are different. This mere logical possibility – which would probably involve evil demons manipulating us, for instance – shouldn't bear on the probability that our actual-world belief is true. But even if we accept this, other kinds of inevitability might still be epistemically relevant. For instance, consider a variant of the case where it is *significantly likely, given the laws of our world*, that our belief about the defendant would remain the same, even when the facts about their guilt are changed.[19] This kind of inevitability just reveals that we didn't have overwhelming evidence to begin with. And when we learn about this inevitability, we should think that the belief in question is less probable than before to be true. It thus seems that the inevitability of a belief must at least be nomologically possible, for it to be able to impact that belief's probability of truth.

The same goes with chanciness. Consider an instance of chanciness that is merely logically possible – it is logically possible, for example, that an evil demon messes with the neurons in my brain such that I form a different belief about p, even when the fact about p remains the same. This kind of chanciness shouldn't impact the probability that my actual belief that p is true. Instead, chanciness must at least be of the nomologically possible variety in order to produce the required epistemic impact.

These restrictions on what counts as an epistemic flaw are consonant with the theoretical argument from earlier. In order for chanciness and inevitability to count as epistemic flaws, they must bear on the truth-conduciveness of the process that we used to form our beliefs – this process, of course, occurs in the actual world. So the kinds of chanciness and inevitability

[18] White actually uses this example to argue that an inevitable belief could still be *justified* – but I believe it could equally well be used to argue that inevitability does not impact probability of truth.
[19] This involves a 'closer' possible world than the one considered previously, because this world is one that has the same laws of nature as the actual world.

that can count as flaws must be the kinds that are possible given the laws of our actual world.[20] Cases like Defendant's Guilt, which invoke merely logically possible inevitability or chanciness, are epistemically irrelevant.

In this part of the chapter, I've argued that various counterexamples don't disable the *Epistemic Premise* of a debunking argument. Cases like Atomic Clock are about the justification of the target moral beliefs, or their status as knowledge – but these epistemic standards are inappropriate for use in moral methodology. We should instead think in terms of estimating the probability of truth of our moral beliefs – hence the debunkers' cases, and their versions of *Epistemic Premise*, are more appropriate. Cases like Defendant's Guilt only show that some kinds of chanciness and inevitability – merely logically possible kinds – are not epistemic flaws. In contrast, nomologically possible chanciness and inevitability could impact the probability of truth of the target moral beliefs. I conclude that debunking arguments are in principle viable, despite these counterexamples.

2.4 Against the Mere Denial of Epistemic Premise

Now step back from the counterexamples and focus instead on the general strategy of denying *Epistemic Premise*. In the rest of this chapter, I argue that we cannot just deny this premise without proposing a replacement – because if we do, it becomes difficult to explain why some evidently problematic beliefs are indeed epistemically problematic.

To start, first notice how I, and most of the literature, have so far viewed debunking arguments as attacks from hostile interlocutors, the debunkers, who are looking to undermine their opponents' moral beliefs. Call this the *adversarial view* of debunking arguments. Adopting this adversarial view naturally lends itself to responses like denying

[20] Even within the nomologically possible worlds, some kinds of chanciness and inevitability might not impact the probability of truth. I explore this in Chapter 4. Moreover, different kinds of chanciness and inevitability might have different impacts on probability of truth – depending, for instance, on how 'close' the relevant scenario is to the actual world. I leave the study of this differential impact for future work.

Epistemic Premise – the response is a purely negative one, aimed at defeating the debunker in rational debate.

But that's not the only way of seeing such arguments. We could instead view them as revealing a conflict in our own belief system – call this the *personal view* of debunking. This conflict arises as follows: we believe some plausible premises – the *Empirical Evidence*, which seems undeniable; *Evidence Reveals Flaw*, which seems plausible too; and, most importantly for this chapter, some form of *Epistemic Premise*. *Epistemic Premise* is especially plausible when supported by the belief pill cases from earlier. Put together, the plausible premises of a debunking argument generate a skeptical threat against the target moral beliefs. An opponent of debunking wants to defend against this skeptical threat, yet they are drawn to it by plausible starting points. Notice that there is no debunker to be defeated here – only conflicting beliefs in a system, and the challenge to explain how we could rationally hold them all together.[21] I think we should adopt this personal view of debunking over the adversarial view, because it better represents the philosophical context: we don't just want to fend off attacks in rational debate – we also want to figure out what to believe about moral matters, and this involves working out a plausible and coherent belief system.

When we adopt the personal view of debunking, the mere denial of *Epistemic Premise* becomes much less attractive. Because if we rejected this premise without proposing any replacements, we become poorly placed to explain why the belief pill cases are so problematic. This pushes us toward an overly permissive attitude toward our beliefs, which seems to be an unacceptable way of resolving the conflict in our belief system. To properly resolve this conflict, an opponent of debunking needs an *Epistemic Premise* that undermines the evidently problematic cases, but doesn't also generate a skeptical threat undermining the target moral beliefs.

[21] Srinivasan (2015, p. 346) also presents this personal view of debunking, and uses it to rebut the objection that debunking arguments are self-defeating. See also Kumar and Campbell (2012, pp. 315–319) and Rini (2017) for similar thoughts, and Enoch (2006, pp. 182–185) and Wright (1991, p. 89) for the same points in different contexts.

My preferred route out of this conflict involves just accepting the versions of *Epistemic Premise* from the previous section – thus granting the in-principle viability of debunking arguments – while denying another premise, *Evidence Reveals Flaw*, in some debunking arguments (see my Chapter 4). This position is similar to McGrath's (2014, p. 213) – she suggests that "if certain possible empirical discoveries were made about why our moral convictions strike us as true, then we should reduce confidence in those convictions. But they [the opponents of debunking] should insist that such evidence actually be provided, as opposed to merely gestured at."

However, there might be other ways out – an opponent of debunking might try to restrict *Epistemic Premise* to cover only the evidently problematic cases (like the ones involving belief pills), while leaving the target moral beliefs unscathed. This limits the in-principle scope of the debunking argument. In what follows, I will explain three strategies for doing so, and argue that they are implausible.

2.5 Three Strategies for Limiting the In-Principle Scope of Debunking

2.5.1 Actual vs Counterfactual Unreliability

The first strategy relies on the distinction between actual and counterfactual reliability. Recall that a belief is actually unreliable when it was produced by a process that has actually produced false outputs; on the other hand, a belief is merely counterfactually unreliable when it was produced by a process that *could have* produced false outputs, but didn't actually do so. An opponent might then try to argue that all the evidently problematic cases (like the beliefs produced by the belief pills) are actually unreliable, whereas the moral beliefs targeted by the debunking argument are merely counterfactually unreliable.

I believe that this strategy won't work, simply because we can't be sure if any of our moral beliefs are actually true or not. So we can't tell whether the process responsible for the target beliefs merely could have produced false outputs before, or whether it actually has produced

false outputs. We are unable to distinguish between actual unreliability and merely counterfactual unreliability when it comes to our moral beliefs – so this distinction cannot help the opponent of debunking.

2.5.2 Non-Moral vs Moral Beliefs

Secondly, an opponent might argue that chanciness and inevitability are only epistemic flaws when they manifest in *non-moral* beliefs, whereas they don't count as flaws when they manifest in moral beliefs. Dworkin (1996) takes this position, and gives variety of arguments in support of it. He argues that taking inevitability to be an epistemic flaw would undermine all our moral beliefs and hence beg the question against a moral non-skeptic; that morality doesn't involve any causal claims and hence our moral beliefs cannot be undermined by discoveries about their causal origins; and that we need to make strong moral assumptions in order to establish a normative connection between our moral beliefs and claims about their causal origins (Dworkin, 1996, pp. 119–127). I believe there are good replies to each of these arguments,[22] but tackling them in detail will take us too far afield. Instead, I'll just draw out a highly unattractive consequence of limiting the scope of debunking in this way – this would render us unable to criticise any moral beliefs on the basis of their chanciness or inevitability, pushing us toward an overly permissive position with regards to our moral beliefs. If we didn't recognise chanciness and inevitability as epistemic flaws, then we wouldn't be able to say that beliefs produced by moral versions of the belief pill cases should have their probability of truth reduced.[23] Moreover, many mundane moral beliefs seem to have their

[22] Here are some brief responses: first, we might argue that inevitability comes in degrees, and that only beliefs that are inevitable beyond some threshold are undermined. Second, when forming our moral beliefs, we make some assumptions about their reliability. Information about causal origins of these beliefs can refute such assumptions, hence undercutting our moral beliefs (McGrath, 2014, pp. 204–211). Third, if we assume cognitivism, and the falsity of a simple subjectivism (on which the moral fact p obtains iff I believe that p), our moral beliefs could come apart from the moral facts, and hence debunking might be possible.

[23] See also McGrath's (2014, pp. 208–209) extension of an example from Dworkin, and White's (2010, p. 598) Coin in the Head example.

probability of truth reduced precisely because of these flaws. Consider, for instance, the following two cases:

> **Doting Dudley:** Dudley loves his son very much. Whatever his son does, Dudley would always believe that his son's action is morally permissible. His son performs an action, and Dudley forms the belief that it is morally permissible.[24]
>
> **Chancy Charlene:** Charlene is reasoning about whether eating factory-farmed meat is morally permissible or not. She thinks about the matter, performs some reasoning, and forms the belief that eating factory-farmed meat is morally impermissible. However, she is careless with her reasoning – sometimes she recognises one subset of the morally relevant factors in her reasoning, and at other times she recognises another subset. This makes her reasoning process quite chancy – she could easily have reached the conclusion that eating factory-farmed meat is morally permissible instead.

One natural way of diagnosing the problem with Dudley's belief is that it is inevitable – had his son performed an action that wasn't in fact morally permissible, Dudley would still have believed that it was. And it is natural to criticise Charlene's belief for being chancy – she could easily have recognised a different subset of morally relevant factors, and reasoned to the opposite conclusion. Upon learning about the chanciness and inevitability in their beliefs, we should lower the probability that Dudley and Charlene's beliefs are true. But if we restricted *Epistemic Premise* to cover only non-moral beliefs, as this second strategy dictates, we would be unable to do so. Thus we can see how *Epistemic Premise* is not only involved in global debunking arguments which undermine all our moral beliefs – it's also required to criticise local, mundane beliefs like Dudley's and Charlene's.[25] Restricting this premise to cover only non-moral beliefs pushes us toward an 'anything goes' approach towards our moral beliefs –

[24] Doting Dudley is based on the example in Kahane (2011, p. 106). The only difference is that Dudley's beliefs are about what's morally permissible, rather than what's admirable.

[25] This point is recognized in different ways by Kahane (2011, p. 115), Huemer (2008, pp. 378–382), and Kagan (2001, p. 54).

which seems an equally unattractive position. To put the point in a different way, Goldilocks' principle applies to our endorsement of *Epistemic Premise*: we want to avoid too much epistemic failure, as with the skepticism foisted upon us by a debunking argument (particularly global debunking arguments which aim to undermine all our moral beliefs). But we also don't want too little epistemic failure, as with being unable to undermine evidently problematic moral beliefs – and some form of *Epistemic Premise* seems crucial for guarding against this.

2.5.3 Proximately Flawed vs Ultimately Flawed Beliefs

The third strategy of restricting *Epistemic Premise* is a little more complicated. It distinguishes between (what I will call) proximate and ultimate flaws. Proximate flaws are ones that arise due to the proximate causes of our moral beliefs – which, recall, are the causes operating in our lifetime that produce our moral beliefs and intuitions. Ultimate flaws, on the other hand, arise from the ultimate causes of our moral beliefs – causes that operate outside our lifetimes, like natural selection over many generations.

To illustrate the difference, I'll first distinguish between proximate and ultimate chanciness. Recall that a belief is chancy when this belief could easily have been different, given different belief-forming circumstances – even when the facts (that the belief is about) remain the same. We can divide this into proximate and ultimate versions. A belief is proximately chancy if it could easily have been different, given some small change in the belief-forming circumstances *within our lifetime*. A belief is ultimately chancy if it could easily have been different, given some small change in belief-forming circumstances *outside our lifetime*. I'll illustrate these using more variations on the belief pill case. In these variations, there are three kinds of belief pills:

- the **believe-p pill**, which makes you believe p
- the **believe-not-p pill**, which makes you believe not-p
- the **random pill**, which creates some chanciness in your psychological states, such that you could believe p, or you could believe not-p

Suppose a doctor administers a belief pill just before birth. Here's how a belief produced by that pill can be ultimately but not proximately chancy. The doctor picks your pill from a bowl with a 50-50 split between believe-p pills, and believe-not-p pills. Let's say they picked the believe-p pill, and you go on to believe p. Your belief is ultimately chancy because it could easily have been different, given some change in the circumstances outside your lifetime – the doctor could easily have picked a believe-not-p pill, and you would have been inculcated into believe not-p instead. But your belief is *not* proximately chancy, because it's not the case that your belief could easily have been different, given some change in circumstances within your lifetime. Once you were given one of the believe-p or believe-not-p pills, you would believed the respective thing, and never deviated from that within your lifetime.

On the other hand, a belief could be proximately but not ultimately chancy. Suppose *all* the pills in the bowl are random pills, and the doctor picks one to administer. With the random pill, there is chanciness in your psychological states, such that you might believe p, or you might believe not-p (just like Chancy Charlene). Your belief is proximately chancy because it could easily have been different, given some small change in your lifetime (some change in your psychological states). But your belief is not ultimately chancy, because there is no change in circumstances outside of your lifetime that would have led to a different belief – you were always going to take a random pill, given how the bowl was set up.[26]

Now think about proximate and ultimate *inevitability*. An inevitable belief is a belief that doesn't change in response to the facts – had the facts been different, we would still have the same beliefs. We might also distinguish between proximate and ultimate inevitability as follows: A belief is proximately inevitable if had the facts been changed *within your lifetime*, you would still have the same belief. A belief is ultimately inevitable if had the facts been changed *outside your lifetime*, you would still hold that same belief.

[26] Here I will ignore possible issues with determinism, and just assume that taking a different random pill will not change the belief that occurs downstream. All I hope to do here is to illustrate the distinction between proximate and ultimate flaws.

For a belief to be ultimately inevitable, it must also be proximately inevitable. Because if there were ultimate causes outside your lifetime that work to inculcate certain beliefs in you, those causes must work *through* the proximate causes within your lifetime that proximately inculcate the belief. Put in terms of the pill example, in order for the ultimate cause (the doctor's picking of the pills) to lead to inevitable beliefs downstream, it must work through proximate causes that themselves lead to inevitable beliefs (i.e. it must only be non-random pills of one kind – either the believe-p or the believe-not-p pills – that are in the bowl). So there are no beliefs that are ultimately but not proximately inevitable.

But a belief could be proximately but not ultimately inevitable. To see this, suppose now that the doctor doesn't just randomly pick a pill. Instead, before taking the pill, they consult the Oracle – who accurately tells them the truth about whether p or not-p. The doctor then administers the corresponding pill to instil the correct belief in you. The belief produced here is proximately inevitable – because had the fact about p been changed in your lifetime, your belief would have remained the same, since you've already been given one of the non-random pills. But the belief is not ultimately inevitable – since if the facts were changed outside your lifetime, the doctor would have learned of the change through the Oracle, and given you the correct pill that then adjusted your beliefs accordingly.[27]

An opponent of a debunking argument could use this distinction between ultimate and proximate flaws to limit the scope of *Epistemic Premise*. They might argue that only proximate flaws can undermine our moral beliefs, whereas ultimate flaws do not. This strategy is only a partial defence, however: it implies that all evolutionary debunking arguments would fail, since such arguments rely on ultimate flaws. But it still allows that debunking arguments based on proximate flaws – such as the order effects debunking proposed by Liao et al. – could work. This strategy will also allow us to undermine Doting

[27] See Nozick (1981, pp. 194–195) for essentially the same case. He doesn't, however, consider proximate and ultimate chanciness.

Dudley and Chancy Charlene's beliefs – since their beliefs are also subject to proximate, rather than ultimate, flaws.

This third strategy could be supported by two lines of argument – I'll outline each one, and then offer a general response to them. The first stems from fairness and moral responsibility. It says that it is unfair or morally repugnant for the epistemic status of our moral beliefs to depend on our being affected by the right kinds of ultimate causes – like being born into the right family or culture, or having descended from beings that were subject to the right evolutionary pressures (Srinivasan, 2015, pp. 347–349). Whether an agent was affected by the right kinds of ultimate causes is out of their control – so it seems unfair to judge that agent negatively when they are influenced by problematic ultimate causes, and form epistemically flawed beliefs as a result. The second argument draws on the concept of justification. It conceives of epistemic status as something that pertains to how well an agent processes and transforms their inputs on the way to forming their belief.[28] But ultimate causes are events well outside an agent's lifetime, and have nothing to do with how well an agent processes their inputs. So we shouldn't hold an agent epistemically responsible for forming epistemically flawed beliefs as a result of dubious ultimate causes.

In response, I allow that perhaps we shouldn't *hold agents responsible* for forming epistemically flawed beliefs as a result of problematic ultimate causes. But it still seems like we should take information about ultimate causes into account when determining the *probability that their belief is true*.[29] Our moral methodology should take all available information into account when estimating this probability of truth, and information about ultimate causes is no exception. Indeed, ultimate causes can often provide useful information about this probability, even if such causes lie outside our lifetimes. For example, suppose you learn that you grew up in a brainwashing regime that was set up by an evil dictator before you were born. The dictator is an ultimate cause that lies outside your lifetime – but information about

[28] Goldman (1979, pp. 12–13) makes some comments that could support this kind of argument.
[29] Srinivasan (2015, p. 349) makes a similar distinction between being blameworthy for one's beliefs, and doing well with one's beliefs.

the dictator's existence should significantly lower the probability that your beliefs are true. Of course, it's debatable whether the influence of evolution is like that of the brainwashing evil dictator. My point here is merely this: an ultimate cause of a belief could still provide useful information about its probability of truth, and we shouldn't rule out the use of such information just because this cause lies outside our lifetimes.[30]

It is possible, however, that ultimate flaws affect a belief's probability of truth less than proximate flaws do.[31] This might be because the proximate causes are causally 'closer' to the target beliefs than the ultimate causes are, and hence are more of a difference-maker to these beliefs. I will set this issue aside – here, I need only show that ultimate flaws can make some impact on the probability of truth of the target beliefs, which will be enough to block this strategy for an opponent of debunking.

2.6 Conclusion

In this chapter, I sought to defend the *Epistemic Premise* of debunking arguments. First, I argued against the use of prominent counterexamples to this premise. Cases like Atomic Clock involve justification or knowledge, but these epistemic standards are inappropriate for a moral methodology aimed at obtaining true beliefs and avoiding false ones. Other cases, like Defendant's Guilt, involve misunderstandings about what counts as an epistemic flaw – inevitability and chanciness must be nomologically possible in order to bear on the probability of truth of an actual belief. In the course of arguing about these cases, I hope to have clarified the nature of the epistemic undermining that should be involved in the consequent of *Epistemic Premise* (undermining probability of truth), and the kind of the

[30] However, good information about the ultimate causes might be more difficult to obtain in practice. In Chapter 4 I argue that we have poor information about the evolutionary causes of our moral beliefs, and that this severely weakens evolutionary debunking arguments.

[31] For instance, Nozick (1981, pp. 194–195) effectively ranks beliefs in the following way, from epistemically better to worse: beliefs with neither ultimate nor proximate flaws, beliefs with ultimate but not proximate flaws, beliefs with proximate but not ultimate flaws, beliefs with both ultimate and proximate flaws.

epistemic flaws that should figure in its antecedent (nomologically possible chanciness and inevitability).

Second, I argued more generally against the denial of *Epistemic Premise*. When we see the debunking argument as revealing a conflict in our own belief system, rather than as an attack from hostile opponents, the mere denial of this premise becomes much less attractive. If we merely deny *Epistemic Premise* without proposing a replacement, we are pushed toward the equally unattractive position of being overly permissive with clearly problematic beliefs. To avoid this unattractive consequence, an opponent could try to propose replacements for *Epistemic Premise* to limit the in-principle scope of debunking arguments – in particular, to undermine only the problematic cases while leaving the target moral beliefs unscathed. I considered three putative replacements – which involve distinctions between actual and counterfactual reliability, between flaws in moral and non-moral beliefs, and between proximate and ultimate flaws – and argued that they are all implausible. I conclude that debunking arguments are based on plausible epistemic premises, and are in principle viable. If we are to resist such arguments, it will not be through denying their *Epistemic Premise*.

3 Debunking Arguments and the Structure of Support in Moral Epistemology

In the previous chapter, I defended debunking arguments against counterexamples to their *Epistemic Premise*. In this chapter, I continue my defence of such arguments – this time, against three further objections that pertain to the structure of support in moral epistemology. These objections don't target any specific premise of a debunking argument – rather, they raise structural issues concerning how debunking arguments fit into the broader web of beliefs. First, the *regress* objection contends that debunking arguments commit us to a problematic regress, and that this regress disables the debunking conclusion. Second, the *findings redundancy* objection claims that the empirical evidence used by debunking arguments is redundant, and such arguments are really just armchair reasoning in disguise. Third, the *argument redundancy* objection contends that debunking arguments assume what they set out to prove, and so their undermining effect is evidentially redundant.

Before tackling these objections, I'll first introduce the tools I'll be using to analysing them – directed graphs that represent the support relations involved in debunking (section 3.1). I then explain and answer each objection in turn (sections 3.2-3.4).

3.1 Debunking Arguments, Directed Graphs

First recall that debunking arguments can be put in terms of the following schema:

> (*Empirical Evidence* about the causal origins of the target moral beliefs, and how your holding of such beliefs depends on various causes)
>
> (*Evidence Reveals Flaw*) If *Empirical Evidence*, then the target moral beliefs have some epistemic flaw F.
>
> (*Epistemic Premise*) If a belief has flaw F, then it is epistemically undermined.

(*Epistemic Conclusion*) Therefore, the target moral beliefs are epistemically undermined. (from *Empirical Evidence, Evidence Reveals Flaw* and *Epistemic Premise*)

Recall also that the personal view of debunking (which I favour) sees these premises as beliefs that we're already committed to, which then generate a skeptical threat against the target moral beliefs – which we also want to hold on to. This skeptical threat is, in essence, an undermining relation that holds between different beliefs of ours, as represented below:

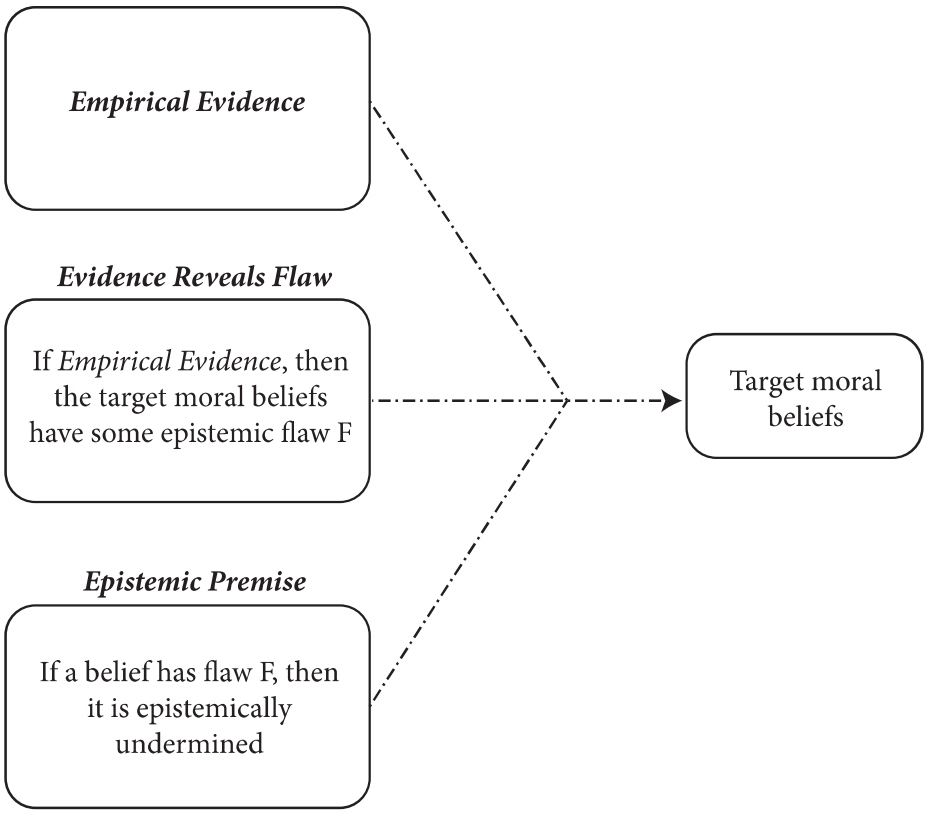


FIG 3.1 The debunking argument as an undermining relation

The boxes represent the moral theorist's beliefs, intuitions, or judgments, and the line represents the relation between them. In particular, the dot-dash line here represents the undermining relation that constitutes the debunking argument – this relation holds between our beliefs in *Empirical Evidence*, *Evidence Reveals Flaw*, and *Epistemic Premise* on the one hand, and the target moral beliefs on the other.

Notice a few important features of this relation. First, it represents an *epistemic* inconsistency, not a logical one. If we held the three beliefs on the left hand side, this reduces the support we have for the target moral beliefs on the right hand side – this makes the target beliefs less probable, or, more generally, gives us an epistemic reason not to hold the target moral beliefs (Sinnott-Armstrong, 2006, p. 223).

Second, the beliefs in the three premises *jointly* undermine the target moral beliefs, rather than doing so independently in an overdetermining fashion. So, for instance, if I only believed *Empirical Evidence* and *Epistemic Premise*, but not *Evidence Reveals Flaw*,[1] these beliefs don't then undermine the target moral beliefs – the undermining effect only obtains when I believe all three premises. This undermining relation is thus a negative instance of what Berker (2015, pp. 330–331) calls 'Y-support', where several beliefs jointly support another belief, rather than doing so independently in an overdetermining fashion.[2]

Third, the undermining relation holds between beliefs of quite different types. *Empirical Evidence* is a non-moral belief about findings from evolutionary or experimental psychology. *Epistemic Premise* also seems non-moral, since it concerns epistemological principles for when a belief – moral or otherwise – is undermined. *Evidence Reveals Flaw* might be partly moral, or might require some moral assumptions, but it is still mostly epistemological, since it's about how the empirical evidence could show our moral beliefs to be flawed in some way. The target beliefs, on the other hand, are clearly about morality. Thus the undermining relation of debunking spans over the web of beliefs, holding across different beliefs that don't have obvious connections with one another. When viewed in this way, debunking arguments introduce new information – specifically, from the empirical evidence, and from epistemology – that could help adjudicate moral disagreements and clashes of intuitions.[3] If a

[1] In Chapter 4, I will argue that this is the case with global, evolutionary debunking arguments from inevitability and chanciness.

[2] Berker (2015, pp. 335–338) uses structures with Y-support to defend coherentism against the circularity and alternative systems objections. I believe the same kind of structure in debunking could help defend against similar objections to moral coherentism too, but I won't pursue that here.

[3] This potential of debunking arguments has been recognized by Singer (2005), Huemer (2008), Sinnott-Armstrong (2006, pp. 230–235), and McPherson (2014).

debunking argument undermines some, but not all, of our moral beliefs – as it does in a local debunking argument – then it can act as a filter for which moral beliefs we should use in our moral theorising.

Representing the debunking argument in this way – with directed graphs – is not just a fancy way of expressing what we already know. As we will see throughout this chapter, these graphs help us visualise the structure of support involved in a debunking argument – highlighting the assumptions it might rely on, and making clear the source of its undermining effect. With a better understanding of this structure, we can then properly evaluate the three objections that are the subject of this chapter.

3.2 The Regress Objection

Start first with the regress objection, due to Regina Rini (2016) and others.[4] This objection goes roughly as follows: to show that some target moral judgments[5] are flawed, a debunking argument must rely on some other moral judgments. But how do we know that these relied-upon judgments aren't also flawed? To check these relied-upon judgments, we might appeal to a further set of judgments. But then we will need to confirm the reliability of this further set too – which relies on yet *another* set of moral judgments! And so this confirmation continues. Rini argues that the defender of debunking arguments is committed to an insurmountable regress, and that this disables the debunking conclusion. If this is correct, then the regress objection poses a problem for debunking. Fortunately, however, I believe that this objection can be overcome. I'll first take a closer look at the objection and illustrate how it works against a specific debunking argument by Liao et al. Then I look at some ways of stopping the regress, using Liao et al.'s argument as a case study. Finally, I argue that even if

[4] See Sidgwick (1907, pp. 212–213) and Berker (2009, n. 76), although I believe these are somewhat different from Rini's objection.

[5] I wish to follow Rini's (2016, p. 679) arguments closely, so I will follow her in using 'judgment' to cover our moral beliefs, intuitions, and any other mental state that could be an indicator of the moral truth. These distinctions will not matter here.

there is a regress, the debunking argument still works. I conclude that the regress objection poses no threat to debunking arguments.

3.2.1 Rini's Regress, and Liao et al.'s Debunking Argument

Rini targets local debunking arguments with her regress objection – recall, these are arguments which hope to undermine some subset of our moral judgments. She cites four such targets: Greene's (2008) debunking of 'characteristically deontological' intuitions, Horowitz's (1998) debunking of intuitions about Quinn's rescue dilemmas, Liao et al.'s (2012) debunking of Loop case intuitions, and de Lazari-Radek and Singer's (2012) debunking of the principle of egoism. In principle, however, I believe this regress objection could affect any debunking argument – so long as the debunking argument relies on some moral assumptions in order to debunk some other moral judgments, it is potentially vulnerable to regress, since it must defend the use of these relied-upon judgments.[6] Thus I believe the regress objection holds deep significance for moral methodology. To make my investigation concrete and tractable, however, I'll examine this objection in relation to Liao et al.'s debunking of Loop case intuitions, which have been found to be subject to order effects. Rather than defending order effects debunking specifically, I hope for this to be a case study in how to defend the assumptions of a debunking argument from a regress objection like Rini's. To that end, while I examine the objection in relation to Liao et al., I will conclude with more general lessons for all moral debunking arguments.

Recall that Liao et al. investigated subjects' intuitions about a thought experiment known as the Loop case, where a trolley is headed toward five innocent people and will kill these five, but could be diverted to a side track to kill one innocent person. This side track loops back to the main track with five innocent people – so if no one was on the side track, the trolley

[6] Thanks to an anonymous reviewer for urging me to clarify this. However, see Rini (2016, pp. 690–694) for dissent.

would loop back and still kill the five (see Fig 3.2). Subjects were asked if it was morally permissible for a person to push a button that would redirect the trolley onto the side track.

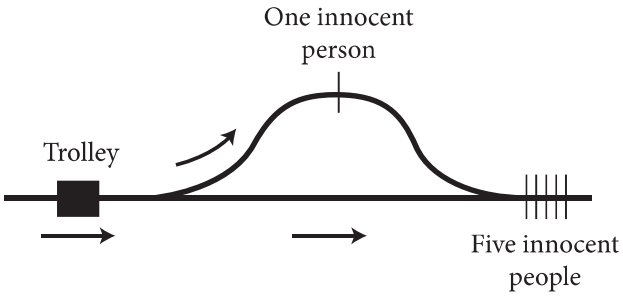


FIG 3.2 THE LOOP CASE

Thomson (1985) famously used this Loop case to argue against the Doctrine of Double Effect, but the details of her argument need not concern us. The interesting issue is this: Liao et al. found that subjects' intuitions about the Loop case varied depending on what case they saw before it (if any). In particular, they tested subjects' intuitions about the Loop case when it was the first case seen, versus when Loop was seen after the Standard case (where there is a side track with one person, but that track doesn't loop back), versus when Loop was seen after the Push case (where, instead of a side track, there is one person standing on a bridge over the track, and subjects are asked if it's permissible to push this one person over, killing them in order to stop the trolley and save the five). Subjects responded to Loop differently, depending on whether they saw Loop first, or whether they saw it after Standard or Push.[7] Given that the moral facts about a case don't change with its order of presentation, Liao et al. conclude that subjects' Loop case intuitions are undermined and should not be used in constructing moral theories. Rini reconstructs their argument as follows:

[7] Liao et al.'s findings actually support two different debunking arguments. The first contends that Loop case intuitions vary with *order of presentation* – for this, the crucial finding is the difference in judgment when subjects see Loop first (before any other case), as opposed to when they see Loop second (after Standard or Push). The second argument contends that Loop case intuitions vary depending on *what case is seen before it* – for that, the crucial finding is the difference observed when subjects see Loop after Standard, versus Loop after Push. Rini focuses on the first kind of argument; so will I.

(*Causal Premise*) Loop case intuitions are caused by a psychological process that is sensitive to order effects.

(*Theoretical Premise*) Processes that are sensitive to order effects do not track the moral truth.

(*Epistemic Conclusion*) Therefore, Loop case intuitions are epistemically undermined. (from *Causal Premise, Theoretical Premise*)[8]

Causal Premise claims that the target judgments about the Loop case are caused by a process with certain properties. *Theoretical Premise* then casts doubt on processes of this kind, claiming that such processes do not track the moral truth. These yield the *Epistemic Conclusion* that the target judgments are epistemically undermined. (Rini's schema is markedly different from the one I've been using, but I've chosen to use it in this chapter, in order to better follow and answer Rini's objection. I believe that our schemas are inter-translatable,[9] however, so nothing important turns on this.)

Before turning to the regress objection, note a terminological point: Rini does not think that debunking argument itself is committed to a regress. Rather, she thinks that in making debunking arguments, the supporter of such arguments – the debunker – will be committed to the regress, and hence cannot support their debunking conclusion.[10] I return to this dialectical point later – for now, we'll focus on the objection itself.

Rini first argues that *Theoretical Premise* needs to be supported by further first-order moral judgments – that is, judgments about the moral properties of specific acts, agents, or consequences (ie. judgments like "Act X is impermissible"). In particular, she thinks that

[8] I omitted a bridge premise, along the lines of "If *Causal Premise* and *Theoretical Premise*, then *Epistemic Conclusion*". I also used the more specific conclusion of the target judgments being "epistemically undermined", because I think this is the most plausible version.

[9] Roughly, the *Causal Premise* of Rini's schema corresponds to my *Empirical Evidence*, and *Theoretical Premise* corresponds to my *Evidence Reveals Flaw* and *Epistemic Premise* (specifically, an *Epistemic Premise* that relies on process-based flaws).

[10] Thanks here to Daniel Stoljar.

Theoretical Premise "is an abstract generalization that gets its support from our unwillingness to accept moral judgments like 'act X is wrong when read-about-first but permissible when read-about-second'." (Rini, 2016, p. 682) That is, Liao et al. can only support *Theoretical Premise* by thinking about specific actions – for instance, harming others for fun, making a false promise, etc. – and considering whether that action could be impermissible when read about first, but permissible when read about second. If we are reluctant to think about actions in this way – and it seems like we are – then we can infer that processes which are subject to order effects do not track the moral truth.

Rini (2016, p. 681) draws an analogy to the task of determining whether a cat's visual system can track spiders. We might check if the cat pounces on spiders at different locations and heights, or spiders with varying amounts of camouflage. We might also test whether the cat pounces on non-spider objects that look like spiders – dust balls, for instance. When testing the cat, it seems like we must rely on our own (human) judgments about which objects are spiders, and which ones are not. In the same way, in demonstrating that a process does not track the moral truth, the debunker needs to rely on their own first-order moral judgments. Rini calls these relied-upon moral judgments the *basis set* – they form the basis for the debunking argument. She thinks that Liao et al.'s basis set must consist of first-order judgments that are reliable.[11]

Next, she argues that a worry arises. The debunker relies on some moral judgments (the basis set) to undermine some other moral judgments (the Loop intuitions). If the target judgments are similar enough to the basis set, then the debunkers have some reason to doubt the basis set itself. Rini draws an analogy to perception: if you learn that some of your perceptions are unreliable, then this gives you reason to worry about other perceptions too. In like fashion, debunking some judgments gives you reason to worry about the basis for debunking, if the target judgments are similar enough to the basis set. For Liao et al., their undermining of the

[11] It is unclear what counts as a reliable basis set. Should that set be free from *all* epistemic flaws, or just the epistemic flaw alleged by the debunking argument? I tackle this issue below.

Loop case intuitions might give them reason to doubt the basis set – that is, the judgments they used to infer *Theoretical Premise* in the first place (Rini, 2016, pp. 682–684, 2016, n. 15).

Rini argues that to be epistemically responsible, the debunkers should confirm the reliability of the basis set. They need to investigate the psychological process that caused the basis set itself, and ascertain that this process tracks the moral truth. To do this, however, the debunker must support a further iteration of claims like *Theoretical Premise* and *Causal Premise* – except now claiming that the basis set was caused by a process that *does* track the moral truth. In Liao et al.'s case, they used some moral judgments to infer *Theoretical Premise* – to be epistemically responsible, they must confirm that such judgments were caused by a process that *does* track the moral truth (Rini, 2016, pp. 683–685).

The confirmation does not stop there, however. To demonstrate that the basis set is reliable, the debunkers must show that the basis set originates from a process that tracks the moral truth. But to show *that*, they need to rely on a yet further set of moral judgments – call this the *further basis set*. Rini argues that the debunker also has reason to doubt this further basis set, if this set is also similar enough to the initial target judgments. Again, we need to confirm the reliability of this further set, and that confirmation requires yet further moral assumptions – so the regress continues (Rini, 2016, pp. 684–685). These steps are only considered abstractly, so it is difficult to say what the further basis set would be in any debunking argument. Still, the general point is clear: to confirm the reliability of the judgments used to infer *Theoretical Premise*, Liao et al. need to rely on some further moral judgments. If these further moral judgments are similar enough to the initial target judgments, however, then these further judgments will also be called into doubt. So the further judgments also need confirmation by some more moral judgments, which will themselves need confirmation, and so on – a regress thus arises.

Rini (2016, p. 685) argues that this regress disables the debunking argument – if the regress continues, the debunkers "never acquire suitable grounds" for supporting the basis set or its further iterations, and so cannot support their argument. Thus, "[i]f the regress does not terminate somewhere, we never reach the debunking conclusion." (Rini, 2016, p. 685)

We can represent and summarise this regress using directed graphs:

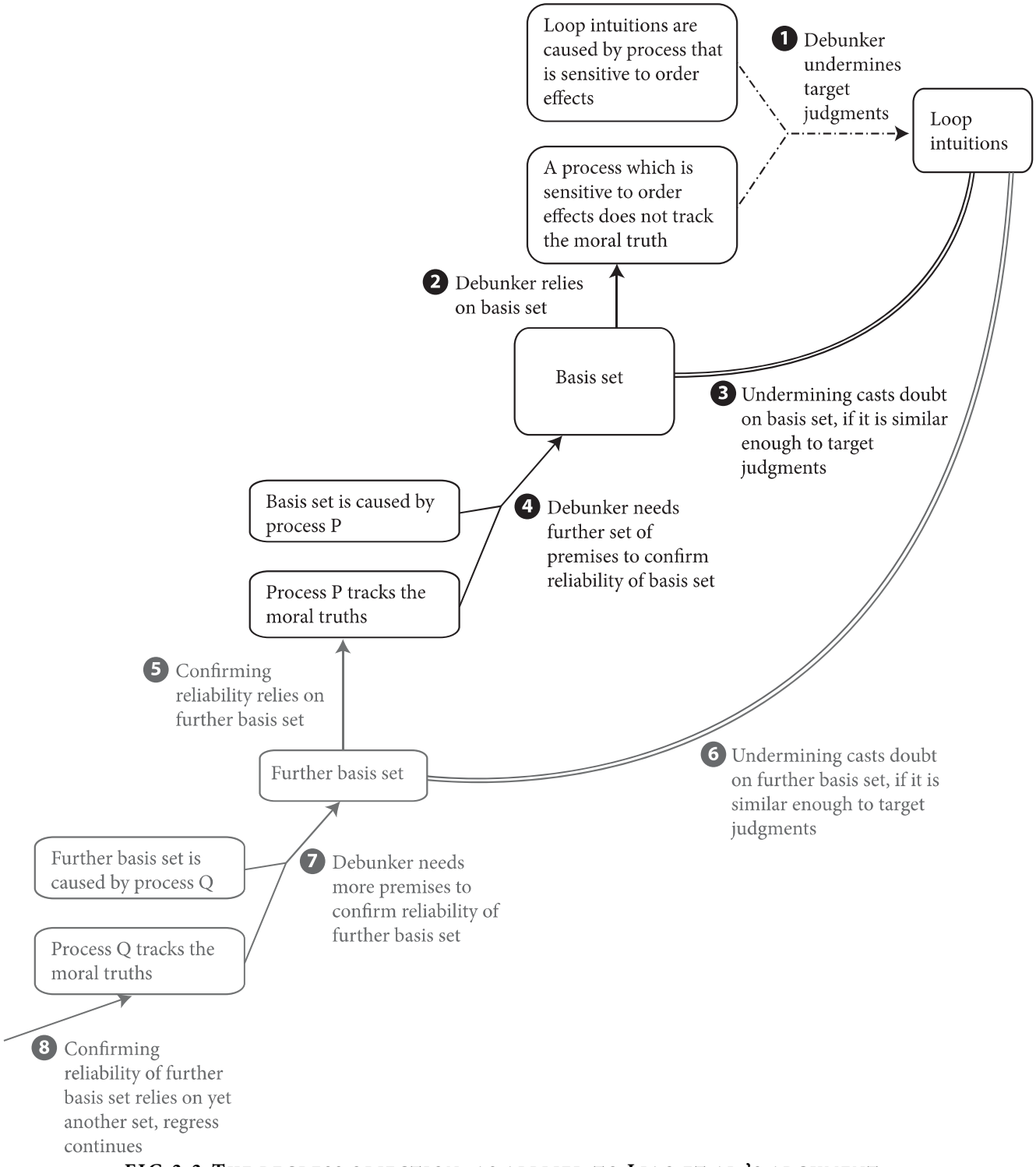


FIG 3.3 THE REGRESS OBJECTION, AS APPLIED TO LIAO ET AL.'S ARGUMENT

As with before, the boxes in Fig 3.3 represent the beliefs, intuitions, or judgments of the moral theorist or debunker. The lines represent relations of positive support, negative support, or similarity – a single solid line with an arrow represents a relation of positive support; a dot-

dash line with an arrow represents a relation of negative support or undermining (which was also presented in section 3.1), and double solid lines represent a relation of similarity that is relevant for epistemic status. Start with the three boxes on the top right. These represent Liao et al.'s initial debunking, where they use *Causal Premise* and *Theoretical Premise* to undermine the target judgments – that is, the Loop case intuitions (step 1 in Fig 3.3). Liao et al.'s argument relies on a basis set to establish *Theoretical Premise* (step 2). If this basis set is similar enough to the target judgments, then they have reason to doubt the basis set itself – this set could be flawed too (step 3). To be epistemically responsible, the debunker needs to rely on a further set of premises, which confirm that the basis set is caused by a process that *does* track moral truths (step 4). But these premises themselves rely on a further basis set (step 5), whose reliability could also be called into question if it is similar enough to the initial target judgments (step 6), and so the regress continues (steps 7, 8). Given this regress, the debunker allegedly cannot support their conclusion.

3.2.2 Resisting the Regress

I'll now look at each step of the regress, and explore how the debunker might respond.

- **Does the debunking argument rely on a basis set of moral judgments?**

Start with whether debunking arguments rely on a basis set of moral judgments. I agree that Liao et al.'s argument relies on a basis set, but this might not be true of all debunking arguments. Some debunkers instead say that to track the moral truths, we must also track the relevant *non-moral* truths – then they take issue with our ability to track these non-moral truths. For instance, O'Neill (2015, pp. 1076–1078) argues that we should not trust moral judgments that have been influenced by sympathy, because sympathy is demonstrably unreliable at tracking when various entities are in pain.[12] In effect, she argues that we fail to

[12] O'Neill (2015, pp. 1076–1078) cites evidence that people attribute less pain to outgroup members (a false negative output, since people don't attribute pain when there is in fact pain), and evidence that people express sympathy for robots being treated violently, in some cases even risking their lives for such robots (a false positive output, where people attribute pain when there is none).

track the moral truths (about whether we should avoid stepping on insects, for instance) because we fail to track the relevant non-moral truths (about whether insects are in pain when we do so). Such debunking arguments don't seem to rely on a basis set of moral judgments, because they are criticising our ability to track non-moral truths. Another kind of debunking argument observes that a process – like emotion-driven processing – fails in some non-moral context, and extrapolates to conclude that it will likely fail in the moral domain too. Such arguments might avoid using a basis set too, although they are quite contentious.[13]

Setting aside these exceptions, I agree that Liao et al. need to make some claim about the nature of moral truth. Without this, it is difficult for a debunking argument to establish anything about the epistemic status of our moral judgments.[14] Now focus on Liao et al.'s argument and their basis set. Recall their *Theoretical Premise*, which says that processes subject to order effects don't track the moral truth. And recall Rini's charge: they must establish *Theoretical Premise* by generalizing from a basis set of first-order moral judgments. They must consider specific actions and think whether they would accept that action as permissible when read about first, and impermissible when read about second. From their reluctance to accept this pair of judgments, they can infer *Theoretical Premise*.

I agree that Liao et al. *could* support their *Theoretical Premise* this way – but that's not the only way of doing so. On an alternative picture, we also have conceptual intuitions about the nature of morality, alongside our first-order moral judgments. When entertaining a claim like "The moral truth about what to do in a case does not depend on that case's order of presentation", it might just appear to be true.[15] Alternatively, we might have the conceptual intuition that "There can be no change in the moral properties described in a case, without a

[13] See Tersman (2008, pp. 392–393), Berker (2009, pp. 316–317) and Kahane (2016, pp. 292–293) for more discussion of this.
[14] Recall Vavova (2014, pp. 92–93) makes a similar point in relation to evolutionary debunking.
[15] Huemer (2008, pp. 383–387) talks about *formal intuitions* – these only impose constraints on moral theories, but do not positively or negatively evaluate anything.

change in the non-moral properties described in that case".[16] If such conceptual intuitions are possible, Liao et al. might appeal to them directly to support *Theoretical Premise*.

Rini (2016, pp. 693–694) anticipates this reply from conceptual intuitions,[17] and argues that "as a purely conceptual matter, it could turn out that the truth of moral judgments is sensitive to their order of presentation", even if we are deeply inclined against such a view. She points to other kinds of truths that might depend on order of presentation – for example, some pairs of counterfactual claims may both seem true when presented in one order, but not so when presented in another order. The order-insensitivity of moral truths isn't a conceptual claim, she concludes – we only arrive at this order-insensitivity by generalizing from first-order moral judgments.

In response, I'll first interrogate the comparison to counterfactual truths, and then offer some positive reasons for thinking that the order-insensitivity of moral truths is indeed a conceptual claim. First, it is not even clear that counterfactual *truths* depend on order of presentation – this is quite a different question from whether our counterfactual *judgments* do. Rini (2016, n. 24) herself admits there is debate about whether the reversibility of counterfactual claims "reflects something deep in the semantics of counterfactuals or is a pragmatic effect." Note too that even if this reversibility bears on the *semantics* of counterfactual judgments, we need a further move to reach conclusions about the *metaphysics* of counterfactual truth. Moreover, even if counterfactual truths do depend on order of presentation, we don't seem licensed to draw any conclusions about moral truths, given the radical difference between the two domains. When I instead consider more similar normative domains – like prudence, epistemology, or even just chess-playing – I struggle to entertain the possibility that truths in these domains depend on their order of presentation to an observer.

[16] This intuition is similar to a formal intuition cited by Huemer (2008, pp. 386–387).

[17] She only anticipates the first conceptual intuition I presented (about the order-insensitivity of moral truth), but not the second one (about the supervenience of the moral on the non-moral). Both, I believe, will support *Theoretical Premise*.

Secondly, there are further reasons to think that the order-insensitivity of moral truths is a conceptual claim about morality, rather than a generalization from first-order moral judgments. Compare the order-insensitivity claim to more typical generalizations like:

(*Pleasure*) All and only pleasure is morally valuable.

Pleasure differs from the order-insensitivity claim in important respects. Our confidence in *Pleasure* varies as we encounter confirming and disconfirming instances of it. For example, when we encounter putative counterexamples like Nozick's (1974, pp. 42–45) experience machine, we tend to reduce confidence in *Pleasure* – or to at least reconsider and take stock of the number of confirming versus disconfirming instances observed. In contrast, when we encounter disconfirming instances of the order-insensitivity claim – that is, when we encounter moral judgments that vary with a case's order of presentation – we just reject these disconfirming instances outright.[18] Importantly, we do not pause to weigh the number of confirming instances against disconfirming ones – this suggests that the order-insensitivity claim is not just a high-confidence claim supported by generalizing from specific instances.

In all, the comparison between moral and counterfactual truths doesn't help Rini's case, and we have further reasons, from our treatment of other moral generalizations, to think that order-insensitivity is part of the concept of morality.

What, then, should we make of Rini's analogy to testing whether a cat can track spiders? To test this, we must seemingly appeal to our own first-order judgments about which objects are spiders. In the same way, we seemingly need to use first-order moral judgments to tell whether a process tracks the moral truths. The analogy is useful, but I believe we have drawn the wrong lessons from it. Because we *can* reach conclusions about the cat's spider-tracking abilities without relying on first-order spider judgments. To see this, suppose I am told that a

[18] Similarly, Huemer (2008, pp. 386–387) argues that the intuition that 'better than' is transitive is not the result of considering specific cases, but rather is produced by our insight into the nature of 'better than' – this explains why we do not immediately accept counterexamples as disproving transitivity, but instead declare such situations "paradoxical".

spider and a dust ball have been placed in separate coloured boxes – a red box and a green box. Each box has a hole cut out at the side – so the cat sitting on the floor can see the item in each box. I cannot see the items, however, so I cannot tell which item is in which box. Suppose I observe, over multiple trials, that the cat pounces on the item in the red box roughly half the time, and on the item in the green box the other half of the time. If the cat always pounces whenever it sees a spider, I can conclude that the cat is not good at tracking spiders. Because regardless of which box the spider is in, I can conclude that the cat correctly identifies the spider only half the time. To reach this conclusion, I need only be sure that the items stay in their respective boxes and never switch places. I don't need information about which item is in which box – which, in effect, would be a first-order spider identification judgment. Analogously, the debunker doesn't need first-order moral judgments to conclude that processes subject to order effects don't track the moral truth. They can get by with something weaker – with just the assumption that the moral truth about a case doesn't change with its order of presentation.

Now step back to the overall debate: Rini wants to argue that Liao et al.'s *Theoretical Premise* can only be supported by generalizing from first-order moral judgments. In contrast, I think conceptual intuitions can also support *Theoretical Premise*. Why does this matter? If Liao et al. can use conceptual intuitions to support their *Theoretical Premise*, they'll have an alternative avenue of support that's different from any first-order moral judgment. To the extent that such alternative avenues are available, the regress objection is weakened. Because such alternatives might be different enough from the target judgments that the debunking argument seeks to undermine – such that the initial undermining does not cast doubt on the basis set. We'll turn now to this issue.

- **Does the initial undermining give us reasons to doubt the basis set?**

Liao et al. want to undermine the Loop case intuitions, and they rely on a basis set to do so. Rini argues that if this basis set is similar enough to the undermined Loop intuitions, then we have reason to doubt the basis set itself – this gets the regress going. However, I believe Liao et al. can resist this.

First, if the previous section is correct, they can rely on conceptual intuitions to support their debunking argument. Conceptual intuitions are at a different *level of generality* from the target Loop case intuitions – so they might be different enough such that the initial undermining doesn't also cast doubt on them. Secondly, even if Liao et al. must use first-order moral judgments in their basis set, this set might still not be similar enough to the Loop case intuitions. Because Liao et al. could use almost any kind of first-order moral judgment to infer their *Theoretical Premise* – many of these judgments will have quite different *content* from the Loop case intuitions. For instance, Liao et al. might consider whether keeping a promise could be permissible when read about first, and impermissible when read about second – and infer *Theoretical Premise* from their reluctance to endorse that pair of judgments. Judgments about promise-keeping have quite different content from the Loop case intuitions: promise-keeping involves a prior speech act by the moral agent, whereas actions in the Loop case do not; promise-keeping need not involve stakes of bodily harm, whereas the Loop case does; and so on. Thus, contrary to Rini, the debunker might find a basis set that is different enough from the Loop case intuitions, such that the initial undermining does not cast doubt on the basis set. These alternatives are illustrated below:

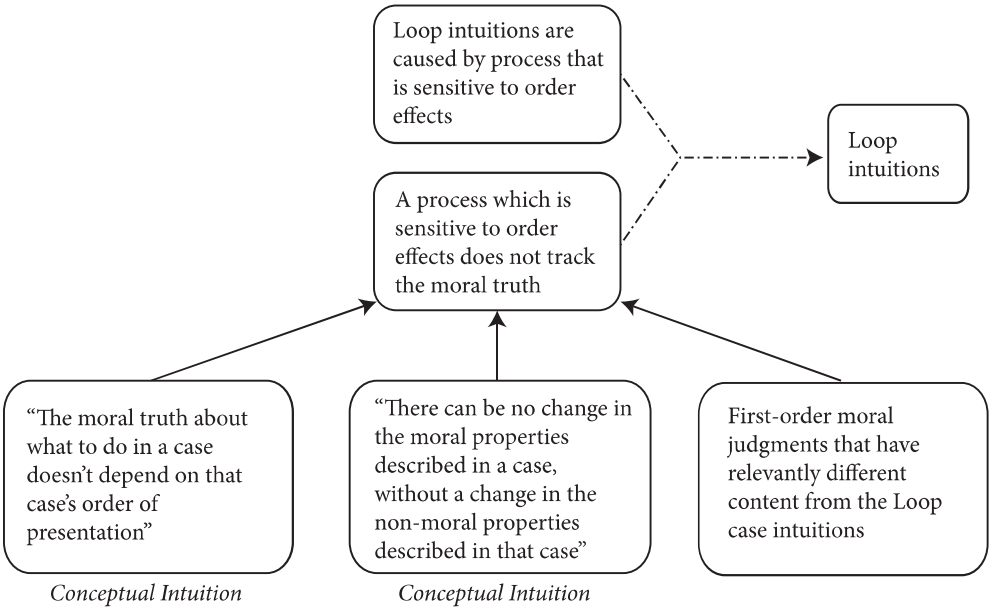


FIG 3.4 ALTERNATIVE WAYS OF SUPPORTING THEORETICAL PREMISE

Rini anticipates the reply that the basis set will be sufficiently different from the target Loop case intuitions. She argues that Liao et al. cannot invoke this reply, because their debunking

argument targets a large set of moral judgments (Rini, 2016, pp. 687–688). Here Rini (2016, n. 15) takes the number of target judgments as a proxy for similarity to the basis set: the more target judgments there are, the more likely there will be a relevant similarity between the target judgments and the basis set – so that undermining the target judgments will also undermine the basis set.

In response, I think we should just reject this strategy of using the number of target judgments as a proxy for similarity.[19] First, because the number of target judgments is a highly imperfect proxy – if the target judgments are of the same kind, for instance, then increasing their number will not increase the likelihood of similarity to the basis set. Secondly, we can directly assess the thing being proxied – similarity to the basis set – so why bother with proxies at all? For instance, I argued above that Liao et al.'s basis set differs from the target Loop case intuitions, because this basis set could have a different level of generality (if they used conceptual intuitions), or because the basis set could have different content (if they used first-order judgments about promise-keeping). In doing so, I have focused directly on the similarities involved, rather than on imperfect proxies. This makes my assessments of similarity – and ultimately, of the epistemic status of the basis set – more fine-grained and likely more accurate than Rini's.

Of course, it is unclear which dimensions of similarity are important for inferring epistemic status. For any particular moral judgment M, this judgment is similar to other moral judgments in various respects. We want to infer M's epistemic status from the epistemic status of these other moral judgments, by virtue of their similarity. To do this, we need to determine which of these other judgments count as 'most similar' for our purposes, which would mean identifying the dimensions of similarity that provide good indications of a judgment's epistemic status. This is difficult, but we might still offer arguments here. Huemer (2008, pp. 383–384) argues, for instance, that the level of generality of a moral judgment

[19] Also see more general remarks by Vavova (2014, p. 98), who notes that "what matters in determining the strength of a debunking argument, or the undermining evidence it provides, is not how many of your beliefs it calls into question but whether it leaves you enough of the right sorts of beliefs with which to evaluate the evidence that has been put before you."

matters: concrete judgments about specific cases are more likely susceptible to emotional bias, and biological and cultural programming, and are hence more likely to be mistaken. Abstract intuitions, on the other hand, are more likely to be the products of rational reflection.

Remember that the whole point of assessing similarity between the target judgments and the basis set is to see whether we have reasons to doubt the basis set. But note a crucial ambiguity here: 'reasons to doubt to the basis set' could either mean reasons to think that the basis set is *flawed in the same way* as the target Loop case intuitions (that is, reasons to think that the basis set might also be subject to order effects), or it could mean reasons to think that the basis set is *flawed in some unspecified way* that might not relate to order effects at all.[20]

If we take the first reading of 'reasons to doubt', then there is a clear way out of the regress for Liao et al. – since their basis set can just comprise of moral judgments *that aren't subject to order effects*. This results in the structure of support shown in Fig 3.5 – the basis set plays a role in supporting itself, when combined with empirical results showing that it was produced by a process that isn't subject to order effects.

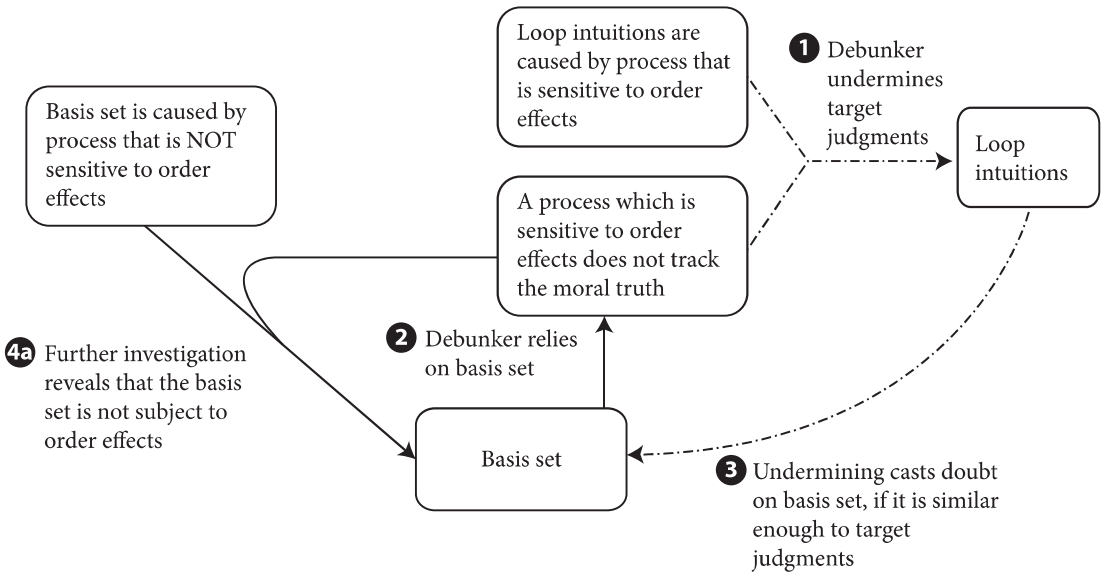


FIG 3.5 SUPPORTING THE BASIS SET WITHOUT RELYING ON A FURTHER BASIS SET

[20] I believe Rini can be read either way on this. For instance, she argues that "the order effects discussed by Liao and colleagues do not seem to be limited in any way; there is nothing in their investigation suggesting that order effects arise only in the particular cases they test." (Rini, 2016, p. 688) This suggests the first reading – that the basis set might be flawed in the same way as the target judgments. But she also talks about ensuring that the basis set is produced by a process that "*does* track the moral truth" (Rini, 2016, p. 684), which suggests the second reading – that the basis set might just be flawed in some unspecified way.

This might be what Rini (2016, pp. 689–690) calls a "self-approving process". She argues, however, that "at the moment we have absolutely no evidence for such a thing." (Rini, 2016, p. 690) This isn't right, at least for Liao et al.'s debunking argument. Because we have found moral judgments that aren't subject to order effects – for instance, Wiegmann et al. (2012, p. 816) find that intuitions about the Push case don't change with their order of presentation. With these results, Liao et al. could use Push case intuitions in their basis set – these judgments been *proven* not to be flawed in the same way as the target Loop case intuitions. Notice that in this case, we can dispense entirely with assessments of similarity to the target judgments. We only wanted to investigate this similarity in order to get more information about the epistemic flaws of the basis set. With the results above, however, we have independently confirmed that the basis set is not flawed in the same way as the target judgments – hence rendering the assessments of similarity unnecessary.

If we take the second reading of 'reasons to doubt', then the standard to meet is much higher. Not only do the debunkers have to show that the basis set is not flawed in the same way as the target judgments, they must show that the basis set *is not flawed at all*. This reading of 'reasons to doubt' is problematic in several ways. Firstly, it is not supported by the empirical, targeted nature of the regress objection. We do not need empirical results or assessments of similarity between different judgments to tell us that our moral judgments are fallible and could be flawed in various ways. If that is all the regress objection amounted to, then it seems to just press a generic worry from armchair skepticism.[21] Compare again with the perceptual case: if I realise that some of my perceptions are unreliable in some specific way, this only gives me reason to worry that similar perceptions will be unreliable in similar ways. It doesn't license me to throw out all my perceptions entirely. So I believe the regress objection is better read in the more targeted manner from earlier: if the target judgments are flawed in some

[21] For a related point about evolutionary debunking arguments, see Vavova (2015, pp. 105–106).

specific way, and such judgments are relevantly similar to the basis set, then we should worry that the basis set is also flawed in the same way.

Secondly, this stronger reading of 'reasons to doubt' is impossibly demanding. We don't demand of others that they eliminate all suspicion of error before considering their arguments. In the same way, we might be holding the debunkers to an overly demanding standard – beyond the scope of proper epistemic responsibility – when asking that they give conclusive proof that their basis set is free from all epistemic flaws.

In this step of the regress objection, Rini argued that the similarity between the targeted Loop case intuitions and the basis set would give us reasons to doubt the basis set. In response, I've argued that Liao et al.'s basis set is relevantly different from the targeted Loop case intuitions. If the basis set consists of conceptual intuitions, then that set has a different level of generality; if the basis set consisted of first-order judgments, then such judgments might have different content. I then considered a response from Rini, who used the number of target judgments as a proxy for similarity. I argued that we should just focus on assessing the similarity directly, rather than on imperfect proxies like the number of target judgments. Finally, I clarified what it means for this similarity to give us reasons to doubt the basis set – I argued that these should only be reasons to think that the basis set is *flawed in the same way* as the targeted Loop case intuitions, rather than reasons to think the set is flawed in some unspecified way. If we take this reading of 'reasons to doubt', Liao et al. can easily evade the objection, since we have already found some moral judgments – those about the Push case – which are not subject to order effects, and would make good candidates for their basis set.

- **Does the initial undermining create further iterations of doubt?**

Finally, quite separate from all the previous issues, notice that the target judgments do all the work in generating the reasons to doubt each iteration of the basis set (see steps 3 and 6 in Fig 3.3). The debunker might argue that at some point, one of the further basis sets is just going to be too different from the target judgments to continue the regress. It's difficult to assess this strategy, because it is unclear what judgments are in the further basis set. But as long as

the notion of similarity used does not deem the target judgments similar enough to *all further basis sets*, the regress will stop at some point.

Let's now summarise the different ways we might resist the regress: craft the debunking argument so that it doesn't rely on a basis set of moral judgments; use a basis set that's different enough from the target judgments (be it in its level of generality, its content, or some other relevant dimension); argue that the similarity between the target judgments and further iterations of the basis set will give out at some point; and finally, discover empirical results showing that the basis set is not flawed in the same way as the target judgments. These strategies are perfectly general – even if you thought that my specific illustration of them (involving Liao et al.) was implausible, they might still be helpful in defending a debunking argument of your choice against similar issues. Moreover, while each strategy might not be equally effective across different debunking arguments, having several strategies on hand creates a formidable defence against the regress.

3.2.3 Even with a regress, the debunking argument still works

The whole purpose of Rini's regress objection was to disable debunking arguments like Liao et al.'s. In the previous sections, I tackled this objection on its own terms: I argued that a regress doesn't in fact obtain – and, in doing so, I accepted (for the sake of argument) the implicit assumption that a regress obtaining would disable the debunking argument. In this section, I question this very assumption – I argue that even if a debunking argument commits us to a regress, that debunking argument can still work.

- **Flows of Support**

First, if we pay attention to the flows of support, it is likely that the regress objection only dampens the impact of the debunking argument, rather than neutralizing it entirely. To see this, first notice that we can understand the regress objection in terms of neutralizing support: an increase in support for *Theoretical Premise* might initially undermine the target judgments. But the target judgments are similar to the basis set, so the initial undermining

also calls the basis set into doubt – this then reduces support for *Theoretical Premise*. So the initial increase in support for *Theoretical Premise* might ultimately neutralize itself. These flows of support are labelled below.

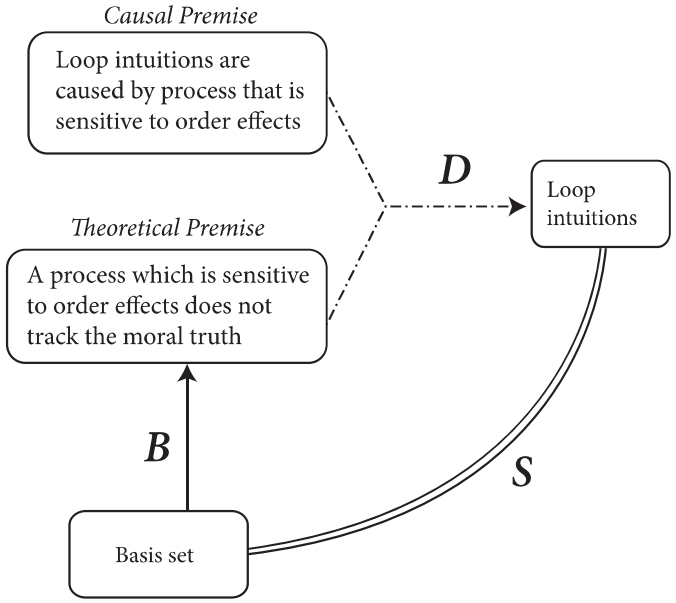


FIG 3.6 ABRIDGED DIAGRAM WITH FLOWS OF SUPPORT

The letters beside each line represent the marginal effect of the relevant relation. For example, D represents the marginal debunking potential of the debunking argument – an x unit increase in support for *Theoretical Premise* should lead to a x*D *decrease* in support for trusting the Loop case intuitions, all other things being equal. (I assume that *Causal Premise* is already strongly supported, so that *Theoretical Premise* is the limiting factor.) D ranges from 0 to 1 – a higher value of D represents a higher debunking potential, and corresponds to a stronger debunking argument. Similarly, S represents the marginal undermining potential that Loop case intuitions have against the basis set of Liao et al.'s debunking argument, by virtue of their similarity. For each x unit decrease in support for the Loop case intuitions, we should decrease our support for the basis set by x*S, other things being equal. Finally, B represents the marginal supporting potential of the basis set in supporting *Theoretical Premise*. Increasing the support for the basis set by x units should lead to x*B *increase* in support for *Theoretical Premise*, other things being equal – and vice versa for decreases in

support. S and B also range from 0 to 1; these values represent the respective strengths of the relevant relations.

Now imagine a 0.1 increase in support for *Theoretical Premise*. If *Causal Premise* is already strongly supported to begin with, then this should translate into a $[0.1 * D]$ *decrease* in support for trusting the Loop case intuitions, through the regular debunking argument. Next, this impact translates into a $[0.1 * D * S]$ decrease in support for trusting the basis set – the similarity between target judgments and basis set means that the basis set is also undermined. Finally, the basis set was supposed to support *Theoretical Premise*, but we reduced support for trusting the basis set, we should now reduce support for *Theoretical Premise* by $[0.1 * D * S * B]$.

We started out with a 0.1 increase in support of *Theoretical Premise*, but this also generated a $[0.1 * D * S * B]$ decrease in support of that same premise. So the overall increase in support for *Theoretical Premise* only amounts to $0.1 - [0.1 * D * S * B]$. For the initial 0.1 increase to fully neutralize itself, we need D, S, and B to all be equal to 1. Remember that D, S, and B represent the marginal debunking potential of the debunking argument, the undermining potential of the similarity between the target judgments and the basis set, and the supporting potential of the basis set respectively. This reveals simple ways for the debunker to avoid the regress objection, when understood as an objection about neutralizing support. The debunker might aim for a weaker debunking argument with a lower debunking potential – one which says, for instance, that increasing the support for *Theoretical Premise* by x units should only lead to a $[0.5 * x]$ decrease in support for the target judgments (debunking potential of 0.5). The debunker might also argue that while there is some similarity between the target judgments and the basis set, this similarity is not complete, so the undermining potential of this similarity is less than 1: each x unit decrease in support for the target judgments should only lead to, say, $[0.7 * x]$ units decrease in support of the basis set (undermining potential of 0.7).[22]

[22] They could also argue that the basis set does not strongly support *Theoretical Premise* – I set that aside here.

In summary, an initial 0.1 increase in support for *Theoretical Premise* will lead to a $0.1 - [0.1 * D * S * B]$ decrease in support for that very premise, generated by the initial regress. If the debunkers can argue, however, that D, S, and B are not equal to 1, then this initial 0.1 increase will not be fully neutralized by the initial regress. Thus the debunking argument will have at least some effect. Notice that this response from the flows of support doesn't contest any premise of the regress objection – it fully accepts that a regress occurs, but allows that the debunking argument can still work. Moreover, note that D, S, and B have a multiplicative relationship in determining the impact of the regress, so a decrease in any of these terms will significantly dampen the overall impact of the regress.

- **Debunking arguments merely highlight an internal inconsistency**

Next recall that the debunking argument itself is not committed to a regress – rather, the supporter of such arguments is committed to the regress. Rini also seems to adopt an adversarial view of debunking arguments – viewing them as hostile attacks put forth by the debunkers, people who have judgments of their own and who could be defeated by a regress. But, as I've argued in the previous chapter, this is not the best way to see the situation. While we often speak of the debunkers as real opponents to be defeated – I myself have done so in this chapter too – there needn't be such opponents at all. Instead, debunking arguments merely highlight an internal tension between *one's own* moral judgments, intuitions, or beliefs – as on the personal view of debunking. Liao et al., for instance, highlight a tension between believing, on the one hand, that moral truths don't depend on a case's order of presentation, and trusting the Loop case intuitions on the other.[23] They can be read as saying that insofar as we're committed to thinking that moral truths don't depend on order of presentation, we shouldn't trust Loop case intuitions. In other words, Fig 3.3 represents our own judgments – not those of the debunker. Regardless of whether there's a regress, the debunking argument –

[23] There might be two other ways to resolve the tension: deny the empirical evidence in *Causal Premise*, or reject the epistemic principles that the argument implicitly relies on. I've suppressed these options because they seem much less attractive.

as illustrated by the three boxes on the top right – still exposes a real tension that needs to be resolved, given the judgments, intuitions, and beliefs that we have.

Of course, merely pointing out an inconsistency is one thing – we need to know how to resolve it (Kumar & Campbell, 2012, p. 318). In Liao et al.'s case, we need to decide whether to keep *Theoretical Premise*, or to trust the Loop case intuitions. Here it seems clear which way to go – we should discard the Loop case intuitions. Because we are much more confident in *Theoretical Premise* – this claim, as seen earlier, is supported by many other moral judgments and intuitions.

Rini is still right to point out, though, that we will never be certain about whether the moral truths are actually sensitive to order of presentation or not. When doing debunking, we are just checking one moral judgment with another, without confirming whether any of them are ultimately accurate. This is a genuine problem, but I think abandoning debunking is the wrong response. Because this is a general problem that affects many other domains – and in these other domains, we still engage in something like debunking. Think, for instance, of the analogous case in sense perception. To check someone's perceptual judgments about whether there are trees or not, I need to use my own judgments about trees. But how do I know if my own perceptual judgments are reliable? I need to check them with some further judgments, which themselves need to be checked, and so on – a similar regress ensues. It does not follow from this, however, that optometrists should not check anyone's eyesight.[24] Even if we cannot confirm that our perceptual judgments are ultimately reliable, there is still epistemic value in trying our best to make these judgments consistent with each other, and with our current empirical findings. The same goes, I think, in the moral case.[25] Saying we should not do

[24] Recall the analogy from Vavova (2015, pp. 105–106), who argues that debunking arguments should be like an optometrist's verdict of colourblindness – the optometrist uses empirical evidence to argue that we're probably making a mistake. This contrasts with the traditional skeptic, who has not produced any empirical evidence, and merely emphasises the possibility of error.

[25] Brink (1989, pp. 129–130) and Sinnott-Armstrong (2006, pp. 243–244) make similar points about coherentism in moral epistemology.

debunking because of the regress objection is analogous to saying that we shouldn't check anyone's eyesight because of the same epistemological regress for sense perception.

To be fair, it might also be that once the relied-upon basis set is made explicit, it's controversial whether this basis set is reliable, so the debunking argument fails.[26] This would be like discovering that my optometrist was drunk when they diagnosed me as colourblind – given their unreliability, it might be rational to resist the debunking and hang on to the target judgments instead. It is difficult to give exact conditions for when a debunking argument succeeds, but here is a tentative proposal: if the basis set is more likely to be true than the target judgments, and the probability that the basis set is true exceeds a certain threshold, then the debunking argument could work. While I believe Rini's regress objection doesn't disable the debunking argument, our investigation into it has drawn attention to this important and neglected requirement: to compare the probability that the basis set is true with the probability that the target judgments are true.

To conclude, I believe the regress objection doesn't disable debunking arguments. I've explored different strategies of resisting the regress, using Liao et al.'s debunking argument as a case study. I then also argued that the debunking argument can still work, even if it commits us to a regress. Firstly, if we pay attention to the flows of support, we can see that the regress objection might only weaken the debunking argument, rather than neutralize it entirely. Secondly, debunking arguments should not be cast as a dialectical attack from opponents with judgments of their own. Instead, such arguments reveal an internal tension within our own judgments: either let go of the basis set, or discard the target judgments. This tension obtains, regardless of whether there is also a regress. And we should, epistemically speaking, resolve the tension. When the basis set is more likely to be true than the target judgments – as I believe is the case with Liao et al.'s argument – we should accept the debunking and throw out the target judgments.

[26] Kumar and Campbell (2012, pp. 317–318) allege this of Greene's debunking argument.

3.3 The Findings Redundancy Objection

At this point, a critic might wonder: if it all comes down to comparing the probability of truth of the basis set with that of the target judgments, why bother with any empirical investigation at all?[27] The allegation here is that the empirical findings seem redundant in a debunking argument. For instance, in Liao et al,'s case, it seems like we only needed to ascertain that moral judgments subject to order effects are unreliable, and this would be enough to undermine the Loop case intuitions – without the need for any experiments or questionnaires. Call this the *findings redundancy objection* – since it contends that the empirical findings are unnecessary for generating a debunking argument.

I'll soon argue that the findings redundancy objection fails, at least for some debunking arguments – but first notice that even if this objection is correct, it doesn't affect the strength of the debunking argument. The objection only alleges that the debunking argument doesn't rely on some premises about empirical evidence that we once thought it did. But this fact, on its own, doesn't dampen the strength of the debunking argument's conclusion – which is solely a function of whether its premises are *true*, and of the strength of the inference from premises to conclusion. Neither of these are affected when we show that an argument doesn't rely on some premises that we thought it did. Nonetheless, if the objection is correct, the debunking argument perhaps isn't as revolutionary as we once thought it was. If the distinctively empirical character of the debunking argument was really just armchair reasoning in disguise, then we can't make as much of the claim that the debunking argument traverses the web of beliefs, making novel epistemic connections between our moral and non-moral beliefs.

Thankfully, though, I think the empirical evidence does play a crucial role, at least in Liao et al.'s debunking argument.[28] Recall that they wanted to undermine Loop case intuitions

[27] Thanks to an anonymous reviewer for suggesting this objection. For similar comments in response to a different argument, see Berker (2009, pp. 325–327) – though he also seems open to the reply I've given below, see Berker (2009, pp. 328–329).

[28] Does the empirical evidence play a crucial role in *evolutionary* debunking arguments, like Joyce's (2006)? Interestingly, both the opponents and proponents of evolutionary debunking seem to think that the evidence

because such intuitions were subject to order effects. For their argument to work, we need *both* the empirical findings (as exemplified in *Causal Premise*) and some theoretical claims about what the moral facts are like (as stated in *Theoretical Premise*).[29] (Recall too that the debunking effect is achieved *jointly* by these premises, and not separately in an overdetermining fashion.) In particular, even if we agree that some class of moral judgments is epistemically flawed (e.g. judgments that are subject to order effects), we still need empirical support for claiming that a specific judgment (e.g. the Loop case intuition) is a member of this class. Thus the overall picture that emerges is this: we agree on some clear epistemic flaws of our moral judgments, and our empirical investigations help identify which specific judgments are flawed in these ways.[30]

In response, a critic might argue that whatever we could glean from the empirical evidence could have just been recognised from the armchair alone. So, again, the debunking argument isn't as revolutionary as we once thought. This might be right for some empirical findings, which merely replicate what philosophers have already discovered. In other cases, however, this seems much less plausible. With the Loop case intuitions, for example, philosophers might not have thought to test for order effects – and even if they did, they might not have detected such effects from the armchair. More generally, when we make a moral judgment about a case, we often only have the judgment that an act is wrong (for instance) – but we don't quite know *why* we judged it to be wrong (Kahane, 2013, pp. 424–428). This judgment might have been influenced by different factors in the case (for instance, to do with whether the trolley track was configured one way or another), or by factors concerning the observer themselves (for instance, what case the observer saw previously).[31] Sophisticated empirical testing can help us determine how these factors causally influence our moral judgments.

---

about evolutionary causes isn't crucial to the evolutionary debunking argument (Vavova, 2015, p. 104; Street, 2006, p. 155; Joyce, 2006, pp. 185–186, 2016, p. 125). I believe, however, that the evidence might be crucial for revealing evolutionary chanciness or contingency, whereas it isn't crucial for revealing inevitability.

[29] Put in terms of my original schema, we need all three premises – *Empirical Evidence*, *Evidence Reveals Flaw*, and *Epistemic Premise* – for the debunking to work.

[30] See Sinnott-Armstrong (2011) for a clear statement of this.

[31] Kahane (2016, n. 44) draws the same distinction, between what he calls epistemically irrelevant influences and morally irrelevant factors.

Then, as moral philosophers, we can think about whether this influence constitutes the tracking of morally relevant features or not. It seems plausible that empirical disciplines will help us identify nonobvious patterns of causal influence in our moral judgments. But to make a truly solid case for the relevance of the empirical findings, we should demonstrate that such findings can uncover patterns that moral philosophers haven't thus far found using their armchairs alone. In Chapter 5, I explore one such pattern, pertaining to how our moral intuitions are influenced by the success probabilities of different courses of action.

3.4 The Argument Redundancy Objection

A different kind of objection alleges that the debunking argument assumes what it sets out to prove – and so its undermining effect is evidentially redundant. Cummins (1998) provides a useful example of this, when criticising the calibration of philosophical intuitions more generally. Suppose we want to check a person's fairness judgments in order to get more information about the fairness facts (assuming there are such facts). The following seems an unhelpful way of doing so: we use a test key that tells us which distributions are fair and which one's aren't, and then check whether the person's judgments align with the test key or not. The problem here is that if we already had the test key, there's no need to check the person's judgments anymore – we have all the information we need about the fairness facts from the test key itself. If we also checked the person's judgments and concluded, for instance, that some fairness judgment of theirs was mistaken, this doesn't give us any further evidence about the fairness facts. In other words, with the test key in hand, this further checking seems evidentially redundant. Cummins (1998, p. 118) concludes that "Once we are in a position to identify artifacts and errors in intuition, philosophy no longer has any use for it." A critic could contend that debunking arguments are also problematic in this way – such arguments implicitly assume that the target judgments are false (like the test key does) and their undermining effect comes entirely from this assumption. Put differently, the debunking argument itself doesn't give us further information about what morality is like – it merely

smuggles in an assumption that target judgments are false to begin with, and it is this assumption, rather than the undermining of the target judgments itself, that's really doing all the work. Call this the *argument redundancy objection*.[32]

To properly evaluate to this objection, it'll be useful to first return to Cummins' example to understand the nature of the redundancy there. Let p be the fairness claim that we want to investigate. The test key tells us that the truth of the matter is p – but, in addition, we also observe someone making a fairness judgment that not-p. The redundancy of checking this fairness judgment can be expressed as follows: all the evidential impact of the test key comes from its providing direct evidence of p – the test key shouldn't also have the *additional* effect of undermining the person's judgment that not-p, and, as a result, further supporting p through this undermining. We can express this redundancy using a directed graph too:

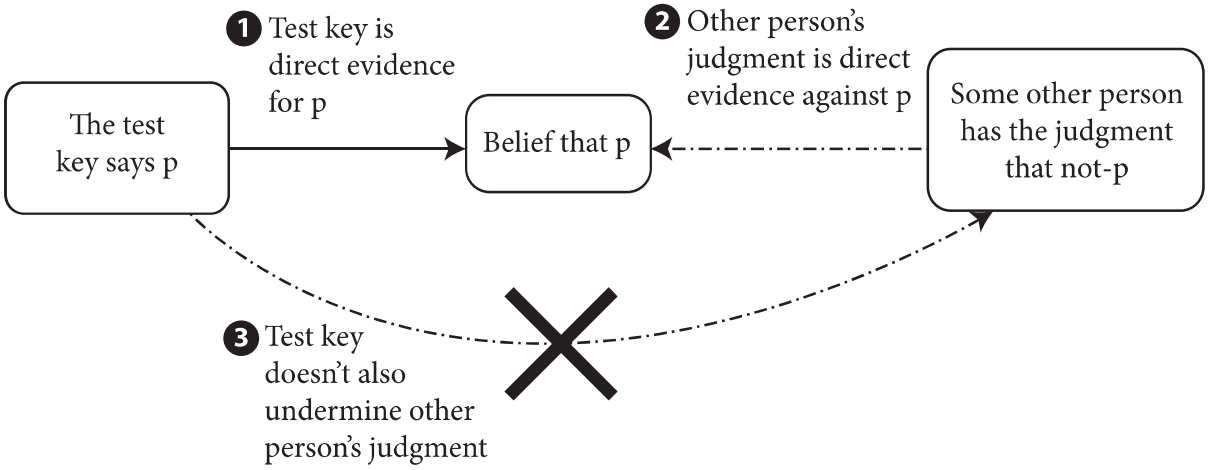


FIG 3.7 REDUNDANCY OF CHECKING THE FAIRNESS JUDGMENT

In the same way, if a debunking argument assumes that p to begin with, and then undermines the judgment that not-p as a result, all its evidential impact stems from whatever independent reason we had for the assumption that p in the first place – rather than from the undermining of the judgment that not-p.

[32] There is much work about whether *evolutionary* debunking arguments beg the question against the moral realist, which I believe is a different issue. But I won't delve into that here – I'll instead just focus on assessing whether local, psychological debunking arguments like Liao et al.'s are evidentially redundant.

I believe, however, that at least some debunking arguments don't fall prey to this objection, because they don't assume the falsity of the target judgments, like the test key does. Think again about Liao et al., who hope to undermine the Loop case intuitions. In their debunking argument, they don't assume any specific verdict about the Loop case – they don't assume, for instance, that we are morally required to turn the trolley to hit the one instead of letting it speed towards the five. If they did, then their debunking argument would be analogous to Cummins's test key case. Instead, they make a weaker assumption: that the moral facts about what to do in a case doesn't change across different orders of presentation. This is still a substantive assumption, but it's weaker than any specific verdict about the Loop case itself. Put in terms of the test key example, their debunking argument is unlike a test key which says that "The answer to Fairness Question 1 is p" – rather, it's more like a test key that says "The answer to Fairness Question 1 is the same as the answer to Fairness Question 2". This imposes some constraints on what the moral facts could be like, but also creates an evidential context where it still makes sense to test or debunk peoples' judgments.

More generally, I believe that as long as a debunking argument doesn't assume that p in order to undermine the judgment that not-p, it isn't evidentially redundant. It might assume something less, for instance that judgments formed in condition A should have the same truth value as judgments formed in condition B (in Liao et al.'s argument, A and B are the different orders of presentation), or that judgments about cases with feature X should have the same truth value as judgments about cases with feature Y. These are substantive assumptions that must be defended with philosophical argumentation, but they don't thereby render the debunking redundant.

As with the findings redundancy objection, I think the best way to rebut the argument redundancy objection is to offer debunking arguments that aren't evidentially redundant. Chapter 5, which focuses on our intuitions about actions with different success probabilities, will contrast debunking arguments that are evidentially redundant with those that aren't.

3.5 Conclusion

In this chapter, I've tackled three further objections to debunking arguments. First was the regress objection, which contends that debunking arguments commit us to a problematic regress, which then disables the debunking conclusion. I argued instead that the regress can be stopped at various steps – for instance, by creating a debunking argument that doesn't rely on moral assumptions, by using a basis for debunking that's sufficiently different from the target judgments, and by drawing on empirical findings showing how the basis for debunking isn't flawed in the same way as the target judgments. I also argued that even if a regress obtains, the debunking argument can still work.

Secondly, I looked at the findings redundancy objection, which alleges that the empirical evidence is redundant in debunking arguments, and that such arguments are really just armchair reasoning in disguise. I responded that sometimes, empirical evidence can be crucial for revealing surprising ways in which different factors influence our moral judgments. Thirdly, I examined the argument redundancy objection, which contends that debunking arguments assume what they set out to prove, and are hence evidentially redundant. I argued instead that at least some debunking arguments aren't redundant in this way – as long as a debunking argument doesn't assume that the target judgments are false, that argument could still have some evidential impact.

4 Debunking Arguments and Evidence of Epistemic Failure

Debunking arguments might be viable in principle, but I believe some of them still fall short because of poor empirical support. In this chapter, I argue that some global, evolutionary debunking arguments are insufficiently supported by the empirical evidence. Evolutionary debunking arguments rely on evidence about the ultimate causes of moral beliefs – which, recall, are causes that stretch far back beyond an organism's lifetime, like natural selection over many generations. Philosophers writing about such arguments tend to focus on the *Empirical Evidence* or on *Epistemic Premise*, with few considering how these premises might interact. In contrast, I hope to fill this gap by explicitly evaluating *Evidence Reveals Flaw* as it appears in these arguments. I argue that the current evidence about the evolutionary origins of our moral beliefs does not show such beliefs to be flawed – roughly, because we are poorly placed to observe, intervene on, and predict what would happen from different evolutionary causes.

This chapter will proceed as follows: section 4.1 details the kind of debunking argument I'll be concerned with in this chapter – global, evolutionary debunking arguments from chanciness and inevitability. Sections 4.2-4.4 investigate arguments from chanciness. I first identify the general conditions for when learning about the chanciness of a belief should lead us to reduce confidence in it. I then argue that these conditions are not met in the case of evolutionary debunking – the current evolutionary evidence doesn't show that our moral beliefs are chancy in a way that warrants reducing confidence. I conclude with general lessons about debunking arguments based on chanciness. Sections 4.5-4.7 treat arguments from inevitability. I also identify conditions for when learning about the inevitability of a belief should lead us to reduce confidence in it, and argue that the current evidence can't establish such inevitability either. I conclude the evidence about the evolutionary origins of our moral beliefs doesn't

establish epistemically problematic chanciness or inevitability – that is, that the respective versions of *Evidence Reveals Flaw* in these debunking arguments are implausible.

4.1 The Target: Global, Evolutionary Debunking Arguments

In this chapter, I will focus on debunking arguments which a) are *global*, in that they target all our moral beliefs at one go; b) are *evolutionary*, in that they rely on evidence about the evolutionary causes of our moral beliefs; c) contend that all our moral beliefs are flawed because they are either chancy or inevitable; and d) conclude that we should lower our confidence in all our moral beliefs (or should think that such beliefs are less probable than before to be true[1]). Such arguments can be represented with the more specific schema below:

> (*Empirical Evidence* about the evolutionary origins of all our moral beliefs)
>
> (*Evidence Reveals Flaw*) If *Empirical Evidence*, then all our moral beliefs are chancy or inevitable.
>
> (*Epistemic Premise*) If a belief is chancy or inevitable, then we should lower our confidence in it.
>
> (*Epistemic Conclusion*) Therefore, we should lower our confidence in all our moral beliefs. (from *Empirical Evidence, Evidence Reveals Flaw* and *Epistemic Premise*)

Joyce's (2016) argument falls squarely into this category. However, what I say below will also be relevant for any evolutionary debunking argument that rely on process-based or disagreement-based flaws – such as Joyce (2006), Handfield (2016), Morton (2016), and Bogardus (2016) – since these also rely on the same version of *Evidence Reveals Flaw*.[2]

[1] I will treat lowering our confidence in a moral belief to be equivalent to reducing the probability of truth of that belief. This follows from the plausible assumption that our confidence in a moral belief should be proportional to its probability of truth.

[2] These other arguments are not of the exact form outlined above, however, because they don't conclude that we should reduce confidence in our moral beliefs. Instead they conclude that our moral beliefs are unjustified or don't count as knowledge.

Recall the flaws that are allegedly revealed by the evolutionary evidence. First, this evidence might show how our moral beliefs are *chancy* in an epistemically problematic way – which involves showing that we could easily have had different moral beliefs (or no moral beliefs at all), given different belief-forming circumstances. An example of a chancy belief was presented in the first belief pill case of Chapter 2: you believe that Napoleon exists, but then learn that you were unknowingly fed one of two pills chosen at random – one pill causes you to believe that Napoleon exists, and another pill causes you to believe that he doesn't. You could easily have been fed the pill that caused you to believe that Napoleon doesn't exist – in which case, you would have very different beliefs, even when the facts about Napoleon's existence remain the same. Learning about this chanciness should lead you to reduce confidence in your belief that Napoleon exists. The evolutionary debunkers might argue that learning about the evolutionary origins of our moral beliefs shows such beliefs to be similarly chancy, hence also concluding that we should reduce confidence in these beliefs.

A related flaw concerns epistemically significant disagreement about our moral beliefs. Recall that Bogardus (2016) argues that the evidence reveals counterfactual evolutionary selves who could easily have existed, and who could have disagreed with us about morality. Disagreement-based flaws are closely related to the flaw of chanciness: if there are nearby counterfactual selves who disagree with us, this indicates that *we* could easily have had different moral beliefs, thus showing our actual moral beliefs to be chancy. Given this close relationship, I will treat the epistemic flaws of chanciness and disagreement together, and just refer to chanciness.

Secondly, the evidence might show how our moral beliefs are *inevitable* in an epistemically problematic way – this involves showing that had the moral *facts* been different, we would still have had the *same beliefs*. The second belief pill case of Chapter 2 was an instance of inevitability: suppose you believe that Napoleon exists, but now you learn for sure that you were slipped a believe-that-Napoleon-exists pill in the past – this pill causes you to form Napoleon beliefs. Had Napoleon *not* existed, the pill would still have ensured your belief that he did exist. Learning about this inevitability should also lead to reduced confidence in your

belief. In the same way, the evolutionary debunkers might argue that evolution has inculcated certain inevitable moral beliefs in us, and that learning about this inevitability should lead to reduced confidence in such beliefs.

But not just any kind of chanciness or inevitability will impact the confidence in our moral beliefs. Recall that in Chapter 2, I argued that chanciness and inevitability must be nomologically possible in order to count as epistemic flaws. There are further conditions for when learning about chanciness or inevitability will mandate reducing confidence in a belief – and I will explore these in the next few sections. I'll first start by examining chanciness.

4.2 How Chanciness Creates a Debunking Argument

Chanciness has the following general form: An agent has the belief that p.[3] They learn that cause A figures into the causal history of their belief. But they also learn that things could have easily been different, such that there was a significant chance that cause B influenced their beliefs differently, leading to their *not believing*[4] that p:

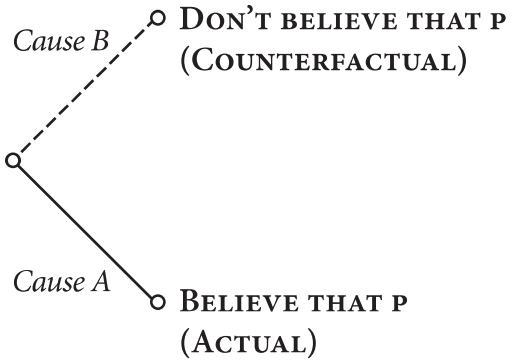


FIG 4.1 Chancy belief that p

As argued in Chapter 2, this does not merely involve the logical possibility that the agent believes something different, as a result of changing the causal history of their belief. Learning

[3] I will assume that believing that p just is having more than 0.5 confidence in p. This choice of threshold level will not matter for my arguments.

[4] Here I take the agent's not believing that p to include their suspending judgment about p, their believing not-p, their adopting a maximally imprecise credence, and so on. These distinctions won't matter for my conclusions.

about this mere possibility only shows that we're nonideal thinkers – which tells us nothing new and hence shouldn't lead to any revision. Instead, chanciness requires some *significant likelihood* that the agent could have believed differently – which could reveal that the agent made a mistake in forming their belief (Christensen, 2007, pp. 208–209; Carey & Matheson, 2013, pp. 139–140; Mogensen, 2016b, p. 601).[5] I won't attempt to spell out this notion of 'significant likelihood' any further, and will just assume for now that the evidence can establish this.

What more does the agent need to learn about causes A and B, so that they should reduce confidence in their actual belief? Drawing on some test cases, I identify some necessary conditions for when the chanciness of a belief mandates reducing confidence in it – that is, I identify the conditions under which chanciness is *epistemically worrisome*. To preview, these conditions will involve the agent's learning more about the counterfactual scenario – about what the facts are like in this scenario, and about their evidence and evidential processing abilities in this scenario.

4.2.2 Necessary conditions on the facts

Start with what the agent needs to learn about the facts. To illustrate, here's a case of chanciness that isn't worrisome. Suppose the agent believes that the chair is blue. They then learn that had the chair been painted red, they would have believed something different – namely, that the chair is *not* blue. This chanciness shouldn't lead them to revise, since the causes in the counterfactual scenario also changed the very truth of the belief in question. Learning of this counterfactual scenario doesn't reveal any flaw in the agent's actual belief, and shouldn't lead to reduced confidence. In fact, learning this might even reveal a *good* feature of the agent's belief, namely that it changes in response to the facts! So learning about this sort of chanciness isn't epistemically worrisome at all.

[5] Compare Bogardus (2016, n. 42).

The lesson: one (often implicit) requirement for chanciness to be worrisome is that in the counterfactual scenario where the agent believes something different – the cause B branch in the figure above – the counterfactual cause doesn't affect the truth of their belief (Ballantyne, 2012, p. 243; Handfield, 2016, pp. 68–70; O'Neill, 2015, pp. 1074–1076).

4.2.3 Necessary conditions on evidence possession

Suppose the agent believes that p. They then learn that had they based their beliefs on a proper subset of the evidence that they actually have, they *wouldn't* believe that p.[6] This chanciness shouldn't lead them to reduce confidence, since it doesn't reveal any rational failing. That they would believe something different in this counterfactual can be attributed to the fact that they had inferior evidence – hence this should not undermine confidence in their actual belief.

In other cases, however, learning of chanciness is worrisome. Suppose the agent believes that p. They then learn that had they phoned an expert about p and based their beliefs on the information obtained, they *wouldn't* believe that p. This seems very worrying, and the agent should reduce their confidence significantly. Because in this case, they learn that had they obtained better evidence and used it, they would have arrived at a different belief.

What about a case where the agent in the counterfactual has *equally good* evidence? This seems worrying too – suppose the agent believes that p, and learns that had they based their beliefs on an equally good, but distinct, set of evidence, they wouldn't have believed that p. This indicates that someone who is in as good an epistemic position as they are believes

[6] This proper subset case appears in different forms in the literature – White (2010, p. 597), for instance, talks about your randomly flipping open a page of a book and forming a belief based on its contents. Here your belief is chancy, but not in a way that is worrisome, because the counterfactual scenario – where you flip to a different page and learn about something else – is one in which you have less evidence concerning your *actual* belief about the page you *actually* flipped to. See also cases of unsafe knowledge – introduced by Comesaña (2005) and cited by Bogardus (2016, pp. 648–650) – and Mogensen's (2016b, pp. 594–595) *Corner Shop* and *My Bayesian Love* cases.

something different – here they should reduce confidence, but perhaps less so than in the previous case.

Revision might be called for in still other cases. Suppose the agent believes that p on the basis of two experiments – studies 1 and 2 – each with a sample size of 100. The agent now learns that had they *only* obtained access to study 3, with sample size of 50 and a completely different set of subjects from the first two studies, they would not have believed that p. If studies 1 and 2 don't provide special information that undermine study 3's results,[7] then it seems like the agent should still reduce confidence here – but perhaps to an even lesser extent than in previous cases. (This follows similar discussion in Christensen (2007, p. 212).)

Aside from appeals to intuition, theoretical principles also justify reducing confidence in these cases. First, the agent could treat their selves in such counterfactual scenarios as reliable testifiers, weighing their opinions in proportion to their epistemic positions.[8] Second, epistemologists have argued that evidence of evidence is itself evidence. In cases where the agent learns of a disagreeing self in a counterfactual scenario, they might have received evidence of evidence against p.[9] In general, what's common in the cases where a reduction in confidence is mandated? The agent learns of a counterfactual scenario where they used *some relevant evidence* to arrive at a different belief.

4.2.4 Necessary conditions on evidential processing

Finally, what does the agent need to learn about their evidential processing abilities in the counterfactual scenario? That is, what do they need to learn about their intellectual faculties (intelligence, memory, perception) and intellectual virtues (thoughtfulness, open-

[7] For instance, studies 1 and 2 might show that experiments with sample size less than 80 are highly unreliable. I leave aside these kinds of interactions, because they don't affect my later conclusions.
[8] See Frances and Matheson (2018, sec. 5.1) for a nice summary of some theoretical principles.
[9] See Feldman (2007, pp. 208–209) for the original presentation in peer disagreement. See also Ballantyne (2014, p. 372), who treats the "evidence of evidence is evidence" (EEE) principle as an analogy, and Mogensen (2016b, p. 600), whose Murder!* case can be given an EEE interpretation.

mindedness, and intellectual courage) in the counterfactual scenario, in order for chanciness to be epistemically worrisome?[10] To illustrate, consider a case from White (2010, p. 599) where chanciness isn't worrisome. Suppose an agent believes that Obama is not an alien. White tells them that five minutes ago, he flipped a coin. Had the coin landed heads, he would have brainwashed them into thinking that Obama *is* an alien. Fortunately for them, however, the coin landed tails and their beliefs were unaffected. Upon learning about this chanciness, should the agent reduce confidence in their actual belief that Obama is not an alien? It doesn't seem so. This is because the brainwashed counterfactual agent has very poor evidential processing abilities – indeed, it is unclear if this agent can process evidence at all. Hence their disagreeing opinions shouldn't undermine the agent's actual beliefs. So another requirement for chanciness to be worrisome is this: the agent also needs to learn that in the counterfactual scenario, their evidential processing abilities are still sufficiently good (such that exposing them to evidence still makes it more likely for them to believe something true rather than something false), and that they used these abilities in forming their belief.

Putting these together, we get some necessary conditions for chanciness to be worrisome:[11]

Consider an agent who believes p. If this agent learns it's significantly likely that a counterfactual cause B would have led them not to believe p, this chanciness gives them reason to reduce confidence *only if* they also learn that:

- B doesn't affect the truth of whether p
- If B were the case, they would have based their belief on some relevant evidence
- If B were the case, they would have used sufficiently good evidential processing abilities in forming their belief[12]

[10] This definition of evidential processing comes from the peer disagreement literature, see Matheson (2015).
[11] I don't need to take a stand on whether these conditions are also *jointly sufficient* – none of my later conclusions turn on that.
[12] Special thanks to Al Hajek, Christian Barry, James Willoughby, and Toby Handfield for help here. These conditions are close to those specified by Ballantyne's (2012, pp. 242–246) symmetry argument, although I don't require that the counterfactual agent has the same total evidence – only that they have some relevant evidence. My conditions are also consistent with Mogensen (2016b), although I go further in adopting a conciliationist position.

These conditions are similar to ones invoked in the peer disagreement literature, which deals with when we should revise our beliefs upon learning about actual people who disagree with us. In the same way, we might think that disagreement with nearby counterfactual selves should lead us to reduce confidence. Indeed, White (2010, pp. 596–597) argues that the impact of debunking sometimes just is the impact of learning about disagreement. Note, however, that disagreement with a counterfactual self might be more worrying than disagreement with an epistemic peer, if your counterfactual selves are more similar to you than mere epistemic peers.[13]

You might also notice that I've come to conclusions resembling 'conciliationist' positions in peer disagreement, which call for reducing confidence in the face of peer disagreement. Conciliationist conclusions can be challenged.[14] But I will assume they are roughly right – since they represent the best prospects for the evolutionary debunker, who wants to argue for an analogous reduction of confidence in our moral beliefs.

So far, we've used simple cases to identify when learning about the chanciness of our moral beliefs should lead us to reduce confidence. Even with such conditions in hand, however, more work needs to be done to apply the conditions to specific controversies (Christensen, 2009, p. 765), such as the one surrounding chanciness and debunking.[15] This is the topic of the next section – where I argue that even if chanciness can undermine our beliefs in the way detailed in this section, the global, evolutionary debunking conclusion still isn't secured.

[13] This is related to Carey and Matheson's (2013, p. 142) point that *agreement* among people with very different histories is more significant than agreement among those with similar histories.
[14] See Frances and Matheson (2018, sec. 5) for an overview.
[15] To my knowledge, only Klenk (2018) and Bogardus (2016, pp. 655–658) have applied peer disagreement principles to evolutionary debunking.

4.3 Chanciness and Evolution

The previous section yielded conditions for when learning about chanciness should lead to reduced confidence in a belief. But are these conditions satisfied in the case of evolutionary debunking? That is, does the following part[16] of *Evidence Reveals Flaw* hold?

> (*Evidence Reveals Chanciness*) If *Empirical Evidence*, then all our moral beliefs are chancy (in an epistemically worrisome way).

I'll argue not. In this section, I'll first outline the global, evolutionary debunking argument from chanciness in more detail. I then argue that its supporters have not provided evidence of epistemically worrisome chanciness.

Recall Joyce's (2006) argument, which could be read as arguing that adaptive explanations of our moral beliefs show such beliefs to be chancy in an epistemically problematic way. More specifically, he can be read as claiming that the belief that "Categorical moral reasons exist" is chancy – that we could easily have not believed in the existence of categorical moral reasons, given different initial evolutionary conditions. Categorical moral reasons (or judgments about such reasons) apply to an agent regardless of what their desires or goals are, and have authority in that they cannot be ignored easily (Joyce, 2006, pp. 57–64).[17] Joyce argues that the tendency to believe in such reasons evolved to solve a weakness-of-will problem in our ancestors, facilitating cooperation and prudent action. We inherited the genetic makeup of our ancestors, which led to our tending to believe in categorical moral reasons (Joyce, 2006, pp. 1-3,107-142). Joyce (2006, p. 181) contends that "[w]ere it not for a certain social ancestry affecting our biology… we wouldn't have concepts like *obligation*, *virtue*, *property*, *desert*, and *fairness* at all." He takes this to undermine the belief that "Categorical moral reasons exist",

[16] For convenience, I will split *Evidence Reveals Flaw* into two parts – one alleging that the evidence reveals problematic chanciness, and another alleging that the evidence reveals problematic inevitability – and treat each one separately.

[17] The folk perhaps don't have the philosophical concept of a categorical reason, but their actions and practices might show tacit commitment to the existence of such reasons. Thanks here to Jessica Isserow.

which then undermines all other moral beliefs too, since all such beliefs entail the existence of categorical moral reasons.[18] In essence, Joyce could be interpreted as making a claim about chanciness, as represented below:

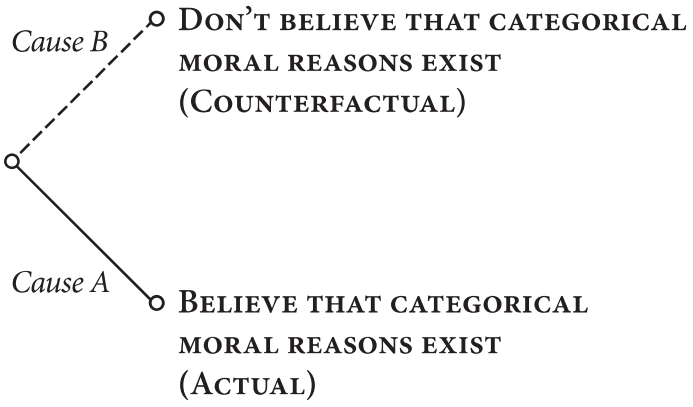


FIG 4.2 Chancy belief that categorical moral reasons exist

Had we evolved under different initial conditions, we wouldn't have believed that categorical moral reasons exist – this, you might think, should lead us to reduce confidence in our actual belief that such reasons do exist.

As it stands, however, these allegations are imprecise. How exactly are the initial evolutionary conditions different in the counterfactual scenario? It's important to get clear on this, because some kinds of evolutionary chanciness seem irrelevant for epistemic evaluation. Consider an agent who learns that an asteroid could have hit the Earth 5000 years ago, leading to extinction of all life and hence to no one believing in categorical moral reasons. It doesn't seem that learning this should mandate reducing confidence in their moral beliefs.

Fortunately, Handfield (2016) has given more substance to this challenge. Taking a similar line, he argues that the tendency to believe in categorical moral reasons evolved due to the *cognitive flexibility* of human beings – that is, due to our ability to revise our ends or goals. Having categorical moral norms was reproductively advantageous, since it motivated our ancestors to cooperate despite many opportunities to defect – opportunities they would easily have taken, due to their cognitive flexibility (Handfield, 2016, pp. 63–67). Thus "[w]e could

[18] As Joyce (2006, p. 181) puts it, "A belief is undermined if one of the concepts figuring in it is undermined." Morton (2016, pp. 245–246) makes a similar move using justificatory closure.

have lacked the belief that some of our norms are inescapable [that is, categorical]… because we could have been much less flexible" (Handfield, 2016, p. 69). This allegation of chanciness can be represented as follows:

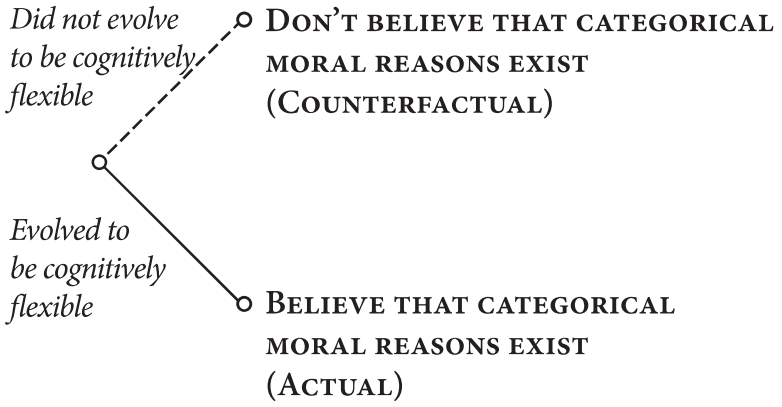


FIG 4.3 Handfield's allegation of chanciness

This refined challenge contends that we could easily have evolved so that we weren't cognitively flexible, and so wouldn't have believed that categorical moral reasons exist.[19] If correct, this might reveal counterfactual selves who believe something different from what we do, generating epistemically worrisome chanciness. Before evaluating the challenge, notice that Handfield imagines quite a radical change to our evolutionary history. It's not clear that the resulting organisms in his counterfactual are identifiably us, or deserve to be called our 'counterfactual selves'.[20] I return to this point later – but for now I'll just use 'counterfactual selves' for convenience.

4.3.2 What our counterfactual selves would believe

Turn now to Handfield's concrete epistemic challenge. We've so far assumed that we learn of chanciness incontrovertibly. But do we actually have evidence of this chanciness, or evidence that it's epistemically worrisome? That is, do we have evidence for claiming that had we

[19] Handfield is interested in counterfactuals where we have different cognitive traits – but the debunkers might instead invoke counterfactuals where the *selective environment* is different from what it is in the actual world. I believe my later arguments will apply equally well to both kinds of counterfactuals. Thanks to Rachael Brown for pointing this out.

[20] Thanks here to Christian Barry and Kim Sterelny.

evolved to be much less cognitively flexible, we wouldn't have believed in categorical moral reasons – or that the three conditions for chanciness to be worrisome are satisfied?

To answer these questions, we need evidence about what would have happened in a counterfactual scenario where the counterfactual cause, B, obtained. There seem to be two main lines of evidence here. First, we might have *actually observed* cases where B obtained – either for ourselves or for someone very similar – and ascertained what happens as a result. Second, we might have a good understanding of the relationship between the causes and our beliefs – such that we could confidently predict what would happen had B obtained, even without actually observing it obtain. In what follows, I argue that the evolutionary debunkers don't have either line of evidence to support their case.

Consider first the evidence for what our counterfactual selves – who are much less cognitively flexible – would believe. Cross-species comparisons will not shed light on what they would believe, since Joyce (2006, pp. 75–85) argues – rightly, I think – that no other species even has the *capacity*[21] to believe in categorical moral reasons. (He takes language to be a prerequisite for belief in categorical moral reasons.[22]) Thus these other species are just not similar enough to humans in order for their situation to shed light on what our counterfactual selves would or wouldn't believe. Comparisons between different human populations will not reveal disagreeing counterfactual selves either – since Joyce (2006, pp. 134–139) takes the tendency to believe in categorical moral reasons to be *universal* across all humans. What the evolutionary debunking argument from chanciness needs is evidence of organisms that have the *capacity* to believe in categorical moral reasons, but do not *tend* to do so because they face different adaptive problems. But we don't have any observational data of that sort.[23] Fraser

[21] Following Fraser (2010, p. 225), I will define the capacity to believe in categorical moral reasons as the ability to have such beliefs, and whatever psychological machinery underlies that ability.

[22] If a debunker holds that language is not a prerequisite for moral judgment, they might still face the problem of determining whether animals without language have moral beliefs. Thanks here to Ole Koksvik.

[23] This argument develops something like Ballantyne's (2014, pp. 377–378) imagined objection 1; it also takes up McGrath's (2014, p. 213) advice to see if debunking evidence can "actually be provided, as opposed to merely gestured at." See also Lewontin (1998) and Kaplan's (2002) pessimism about finding good evidence about the evolution of human cognition, and Isserow's (2018) scepticism about determining the true genealogy of our moral beliefs.

(2010, p. 225) argues against a view on the opposite extreme, which holds that organisms with a requisite level of intelligence will *inevitably* tend to make moral judgments.[24] He argues that "[w]e have no comparably intelligent, social creatures here on Earth to examine. No alien species suitable for testing the claim have yet deigned to visit us." My argument here just poses the same kind of challenge to the evolutionary debunker's claim that our moral beliefs are evolutionarily chancy.

But perhaps the evolutionary debunkers could rely on a different kind of evidence – a good understanding of the relationship between the cause (lack of cognitive flexibility) and the belief (in the existence of categorical moral reasons). This understanding might allow us to predict what we would believe, had we been much less cognitively flexible. But I don't think this strategy is promising either. Joyce (2006, pp. 123–124) himself points out, for instance, that we know very little about the neurological and genetic bases of the tendency to believe in categorical moral reasons – and it seems like only *extensive* knowledge of this kind will help prediction here.

Still, suppose for the sake of argument that the debunkers could establish that had our *ancestors* evolved such that they lacked cognitive flexibility, *they* wouldn't have tended to believe in categorical moral reasons. It doesn't immediately follow that *we,* modern humans, wouldn't have developed this tendency in the counterfactual. It could be that between the time of the ancestral selective environment and the present day, some other causes led us to develop this tendency anyway.[25] This is not just idle speculation about overdetermination – it's a real possibility. For instance, Sterelny (2010) argues that the tendency to use moral concepts stems from developmental causes: parents provide social signals of failure for trial and error learning, they teach their children using toys, tools, stories, games, and examples – these all engage the child's pattern-recognition abilities, which then generate moral

[24] For example, Ayala (2010) argues that the tendency to make moral judgments is a by-product of certain intellectual faculties, while Wielenberg (2014, pp. 170–173) argues that it might nomologically impossible for creatures that evolved like us to have radically different moral beliefs.

[25] Also see O'Neill (2015, pp. 1072–1074) on how an unreliable ultimate cause of moral belief could nonetheless produce a reliable proximate cause.

judgments. If developmental causes like this are in place, then our tendency to make moral judgments would not be evolutionarily chancy either. This highlights how the evolutionary debunking argument from chanciness is committed to a strong counterfactual claim about what we would believe under different initial conditions (see figure below) – one that just isn't well-supported by the current empirical evidence.

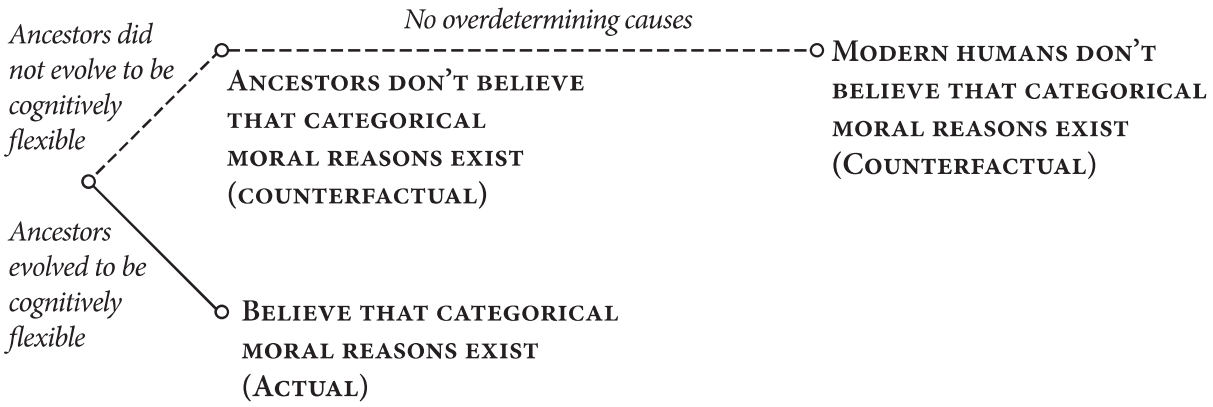


FIG 4.4 Evolutionary chanciness is a strong counterfactual claim

Thus the debunkers don't even have evidence that our counterfactual evolutionary selves would disagree with us. Furthermore, they must also establish that it's *significantly likely* that the evolutionary causes would have been different in the first place, which then leads to the emergence of such counterfactual selves.[26] This poses a further obstacle which I'll set aside for the purposes of this paper. In sum, though, it's unclear if there is even chanciness to begin with.

4.3.3 Moral facts in the counterfactual

What about the three necessary conditions for chanciness to be worrisome? The first condition says that the counterfactual cause (lack of cognitive flexibility) must not affect the truth of the belief in question (that categorical moral reasons exist). Here, the debunkers need to defend a metaphysical claim – namely, that the moral facts that obtain in the

[26] Thanks here to an anonymous referee. I have focused on what we would or wouldn't believe in the counterfactual, rather than on how likely this counterfactual scenario would obtain – both, of course, are relevant to assessments of epistemically worrisome chanciness.

counterfactual scenario are the same as the moral facts that obtain in the actual world.[27] We might push back on this, and argue that cognitively inflexible organisms are so unlike us that no moral reasons (or different moral reasons) will apply to them in the counterfactual. After all, these organisms cannot respond to reasons like we do, and hence might not even be moral agents at all – nor will they face the same challenges concerning cooperation and coordination.[28] Handfield (2016, pp. 70–71) objects, arguing that our cognitive flexibility shouldn't affect the nature of the moral facts (nor should it affect which moral reasons will apply). Here we seem to reach a stalemate: two competing views about the nature of moral facts, and no way to decide between them. I'm not sure what follows: does the debunker satisfy this requirement, because their opponents cannot beg the question by assuming something about the nature of the moral facts? Or is the debunking challenge disabled, because it too relies on such assumptions?[29] I'm not keen on pushing the burden of proof around here, so I'll note this conflict and move on.

4.3.4 Our counterfactual selves' evidence and evidential processing abilities

We still have two further conditions – the debunkers need to show that our counterfactual selves have some relevant evidence, and that such selves have sufficiently good evidential processing abilities. Here the debunking argument clearly falls short. While some evolutionary debunkers liken their arguments to parallel ones from actual peer disagreement (Bogardus, 2016, p. 656; Handfield, 2016, pp. 80–81), there are crucial disanalogies between these two types of argument. Consider Christensen's (2007, p. 193) restaurant check case from the peer disagreement literature: we are in a restaurant and decide to split the bill equally among five people. I do the math in my head and become highly confident that our

[27] Notice that this claim is different from, and significantly weaker than, the claim that the moral cannot be reduced to the non-moral – which is what Joyce (2006, pp. 188–190) hoped to establish.
[28] Thanks here to Kim Sterelny. See Copp (2008), Lillehammer (2010), Sterelny and Fraser (2017) for examples of moral theories which will claim that the moral facts are different in the counterfactual scenario.
[29] See Sinclair (2018, sec. 6.2), Vavova (2014, p. 93), Street (2008, pp. 215–216) and Copp (2008, pp. 195–202) for more discussion of this.

shares are $43 each. My friend does the same, and becomes highly confident that our shares are $45 each. We both have equally good track records of past calculations – we usually agree about the calculations, but when we disagreed, my friend was right as often as I was. Christensen argues that I should become less confident that our shares are $43 – because there is a symmetry in evidence and evidential processing ability between me and my disagreeing friend. The evolutionary debunkers contend that learning about disagreeing counterfactual selves should have the same effect. But notice an important difference – in the restaurant check case, we have information about the *track record* of the disagreeing party, which establishes the required symmetry. The evolutionary debunkers, on the other hand, have not produced any track record of our counterfactual selves' judgments. In fact, they haven't provided any information about other judgments that these counterfactual selves would make,[30] or about the relative epistemic position of such selves.[31] And, for reasons rehearsed earlier, I doubt we'll ever get such information. These differences make the debunking argument significantly weaker than the traditional argument from actual peer disagreement.

In response, the debunkers will insist that they don't need to give positive reasons for thinking that our disagreeing counterfactual selves are in a decent epistemic position. All that's needed for debunking is the lack of positive reason to privilege our beliefs over that of our counterfactual selves (Ballantyne, 2012, pp. 250–253; Bogardus, 2016, pp. 657–658).[32] Against this, I'll first cite an epistemic intuition: it doesn't seem the default position to just assume that anyone who disagrees with you has some relevant evidence – or that they have

[30] Contrast this with cases of actual disagreement about controversial moral issues – like whether it's right to administer the death penalty. Such cases are worrying because the disagreeing party *agrees with us* about many other moral issues – they agree that lying is prima facie wrong, that it's wrong to cause pain for your own amusement, and so on (McGrath, 2008, pp. 103–106).

[31] I'm not demanding that the counterfactual selves meet a strict standard of evidential symmetry, as King (2012) has demanded with actual peer disagreement. I'm pointing out that the debunkers can't even establish a rough evidential symmetry, because they have given no further information – not even a general idea of how we fare on the epistemic spectrum of counterfactual organisms. (Christensen (2007, p. 212) allows that this general information might be enough to mandate reducing confidence.)

[32] This might correspond to Vavova's distinction – outlined in Chapter 1 – between an argument giving us good reason to think our moral beliefs are flawed, and one that demands good reason to think those beliefs aren't flawed.

sufficiently good evidential processing abilities. Think, for instance, of an agent who learns that someone disagrees with them, but learns nothing more about what that person is like.[33] It doesn't seem that learning of this disagreement *alone* should prompt a reduction in confidence. There's also a theoretical consideration against revising here. The disagreeing party could just as likely be an epistemic inferior as an epistemic superior or peer – so reducing confidence in response to their disagreement could just as well lead the agent to epistemically *worse* beliefs. Perhaps it's rationally *permissible* to reduce confidence in these cases, but it's difficult to see why the agent would be rationally *required* to do so – and this requirement is what the debunkers need.

Finally, on Handfield's specific argument, our counterfactual selves – who lack cognitive flexibility and are unable to revise their ends or goals – would be very different from us. Handfield (2016, p. 69) allows that "[p]erhaps the resulting organisms would not be readily recognizable as 'human'." (Contrast with the agents in the counterfactuals of section 4.2 – who, for instance, merely obtained more evidence than we did.) As mentioned earlier, we might think that Handfield's counterfactual organisms aren't 'us' at all. So even if we concede that their disagreement matters, this marked difference from us dampens the impact of such disagreement. Moreover, if morality involves 'stepping back' from one's goals, assessing such goals and revising them,[34] we might even have independent reasons[35] for claiming that such organisms have *poor* evidential processing abilities – since they can't revise their goals at all.

In sum, there are further obstacles to arguing that our counterfactual selves have the relevant evidence, or that they have sufficiently good evidential processing abilities. Thus, in the global evolutionary debunking arguments we have considered, *Evidence Reveals Chanciness* does not hold – thus these arguments do not work, or are at least significantly weakened.

---

[33] Thanks to Christian Barry for spurring me to think about this example. It's an extreme case of what Frances and Matheson (2018, secs. 4, 5.5.4) call *disagreements with unknowns*.

[34] This picture of practical reason is most often associated with Kant. I believe, however, that even those who accept different views will still want to say that practical reason involves being able to revise at least some of one's goals – and the counterfactual selves in Handfield's argument are just unable to do that.

[35] Roughly, an independent reason to discount others' disagreement is a reason that doesn't rely on the disputed belief, or on the reasoning behind that belief (Christensen, 2007, pp. 198–199; Elga, 2007, pp. 489–490).

At this point, the debunker might invoke a different kind of counterfactual organism from the one Handfield proposes. This new counterfactual organism is cognitively flexible – it can revise its goals – but it's also ideally prudentially rational, and never succumbs to the temptation of short-term gains. Such organisms might not believe in categorical moral reasons either, if the tendency to believe in such reasons evolved to counteract our temptation towards short-term gains. I concede that such organisms might be more similar to us, and perhaps they meet the requirement for evidential processing. But still, we just don't know what such organisms would or wouldn't believe – it's just a conjecture that they wouldn't believe in categorical moral reasons, rather than a likely claim supported by good evidence. It might also be that the moral facts are different for such creatures – thus violating the first necessary condition – since again they would face very different kinds of cooperation problems than we do.[36]

Given all my criticisms, it's natural to ask: what evidence *do* the debunkers have to support the claim about chanciness? First, they cite the alleged universality of moral norms across different human societies (Joyce, 2006, pp. 134–135). Secondly, they invoke developmental evidence concerning the early and reliable emergence of the moral/conventional distinction, and the ability to reason with deontic conditionals (Joyce, 2006, pp. 135–136). I need not take a stance on the reliability of this evidence – my arguments above aim to show that *even if* such evidence is reliable, more needs to be done to yield an evolutionary debunking argument from chanciness.

4.4 Lessons about Chanciness

In section 4.2, I examined what kinds of chanciness would mandate reducing confidence in a belief. Then, in section 4.3, I applied this analysis to the global, evolutionary debunking arguments suggested by Joyce, Handfield, and Bogardus. I argued that there is insufficient

[36] Thanks here to Kim Sterelny for suggesting this alternative and some of the responses.

evidence for claiming that chanciness obtains – we don't have evidence for claiming that our counterfactual selves (or, rather, some counterfactual organisms like us) wouldn't believe that categorical moral reasons exist. Moreover, even if we concede that this chanciness obtains, it still doesn't meet the three necessary conditions to be epistemically worrisome. The moral facts might have been different in the counterfactuals they imagine; the resulting counterfactual organisms might not have the requisite evidence or evidential processing abilities for their disagreement to matter. Recall that these are all *necessary* conditions – so failing to satisfy any would disable their challenge. Thus *Evidence Reveals Chanciness* is implausible in the case of these global, evolutionary debunking arguments from chanciness.

It's possible that the debunkers could marshal further evidence to establish their claims, or argue that the relevant evidential standards are already met. At the very least, however, I hope we can agree that when this debunking argument's impact on our moral beliefs depends on how confident we are that our counterfactual selves would disagree with us, and that the three necessary conditions are satisfied. These form dimensions along which debunking arguments from chanciness could be made more or less convincing[37] – and I've argued that these arguments are so far on the 'less convincing' end.

Recognising these dimensions means that my earlier criticisms might not apply with equal force to claims about the *historical* or *cultural* chanciness of our moral beliefs, for instance. The difference between these sorts of chanciness and those discussed earlier is that we might be better placed to observe, intervene on, and predict what would happen had we been exposed to different cultural conditions, for instance – and hence be better placed to establish chanciness and the three further conditions for chanciness to be worrisome. In particular, with *experimental* chanciness, we will be most confident that chanciness obtains, and that the three further conditions are satisfied – this can then generate a plausible debunking argument. Consider an alternative debunking argument based on experimental findings – we randomly divide subjects into two groups, expose them to different causes, and check what

[37] However, variations along these dimensions might not linearly affect the impact on our moral beliefs.

beliefs result. (An example of this is Liao et al.'s (2012) order effects study.) The chanciness of our moral beliefs, if found by the experiment, would be much better established. This, I believe, translates into a better-supported debunking argument from chanciness.

4.5 How Inevitability Creates a Debunking Argument

Let's now turn to the epistemic flaw of inevitability. We might instead learn that evolutionary processes have strongly inculcated certain moral beliefs in us, making such beliefs unacceptably inevitable – so that had the moral *facts* been different, we would still have the same moral *beliefs* about them. Recall the example of inevitability from Chapter 2: you learn that you were previously fed a believe-that-Napoleon-exists pill, which would have ensured your belief that Napoleon exists, regardless of whether he actually did.

Unlike the earlier arguments from chanciness, debunking arguments from inevitability rely on a counterfactual where there's a change in the *moral facts*, rather than a change in the circumstances surrounding belief formation. But I believe arguments from inevitability also suffer from problems with the empirical evidence – I'll now question the other part of *Evidence Reveals Flaw*:

> (*Evidence Reveals Inevitability*) If *Empirical Evidence*, then all our moral beliefs are inevitable (in an epistemically worrisome way).

As with before, let's look at some simple cases to better understand what's required for inevitability to be epistemically worrisome. Inevitability can be understood as follows: An agent has the belief that p. They learn that cause A figures in the causal history of their belief, and that this cause would have made them believe that p, whether or not p was actually true.

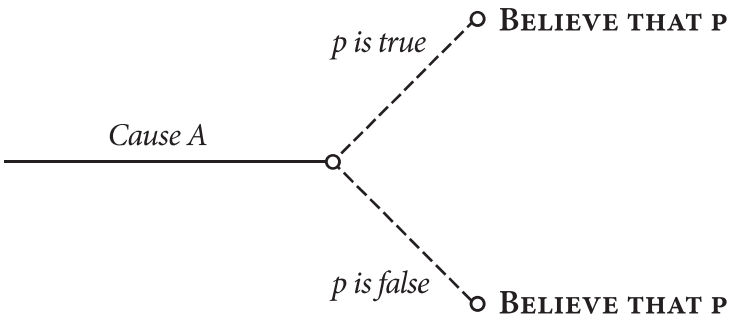


FIG 4.5 Inevitable belief that p

I'll assume that the agent doesn't have information about whether p is actually true or false – that is, they don't have information about which branch they are on – since this seems like a plausible way to model the epistemic situation with our moral beliefs. All the agent learns is that had the truth of p been changed from what it is in the actual world, their belief that p would remain the same.[38] This is epistemically worrisome because it shows that in one of these branches – either the scenario where p is true, or where p is false – they have failed epistemically. They are insensitive or unresponsive to the facts, since they haven't changed their beliefs to match the world.

This is also an intuitive way of modelling the red wall case from Chapter 1: upon learning about the red light shining on the wall, you realise that you would have believed that the wall was red, whether or not it was actually red. Had the colour of the wall been changed – from red to white, or the other way around – you would still have formed the same belief that the wall was red.

Notice that unlike with chanciness, the agent can't treat their selves on either of these branches as reliable testifiers, and can't use evidence of these selves' evidence as first-order evidence. This is because one of these selves inhabits a world where the facts are different than what they are in the actual world – and the agent isn't sure which self is which. Instead, learning about inevitability only provides some higher-order evidence to the agent – they learn that their actual evidence or evidential processing abilities are problematically

[38] This is an important difference between my cases and more standard cases pertaining to whether sensitivity is necessary for knowledge. In these more standard cases, we are always considering a belief that's *true* – whereas in my cases, I make no such assumption.

insensitive or unresponsive to the facts. And, to that extent, they should reduce confidence in beliefs that were based on such evidence and abilities.

Sometimes, learning that a belief is inevitable mandates that we should reduce confidence in it. In other cases, however, it does not. As with chanciness, the challenge is to identify the necessary conditions for inevitability to be epistemically worrisome.

4.5.2 Necessary conditions on evidence

Suppose an agent forms a belief that the wall is red, based on their sensory perception of a red wall. They then learn that had they based their beliefs on an informant's testimony (rather than on their own sense perception), they would have believed that the wall was red, regardless of whether or not it actually was red. Learning this information shouldn't mandate reducing confidence in their actual belief. This is because the information learned pertains to the unresponsiveness of a different source of evidence – testimony rather than sense perception – which wasn't used to form their actual beliefs. So for inevitability to be epistemically worrisome, the agent must learn that their beliefs would be inevitable when based on evidence that's *relevantly similar* to the evidence they used in the actual world. (Contrast this with the red wall example presented in Chapter 1, where presumably we learned that our belief would be inevitable when based on sense perception – which is the same source of evidence we actually used to determine the colour of the wall.)

4.5.3 Necessary conditions on evidential processing abilities

Suppose again that the agent forms a red wall belief using sense perception in normal conditions. They then learn that had they formed their beliefs using sense perception while they were drunk, they would have believed that the wall was red, regardless of whether or not it was. Such information also doesn't reveal any epistemic failure in the formation of their actual-world belief, and shouldn't mandate reducing confidence. In the actual world, they

used sufficiently good processing abilities to form their belief – and these abilities are not impugned by learning about the counterfactual inevitability when they are drunk. So another necessary condition is this: the agent must learn about inevitability in cases where they used sufficiently good evidential processing abilities to form their beliefs.

We thus get some necessary conditions for inevitability to be epistemically worrisome:

Consider an agent who believes p. If this agent learns that they were subject to a cause A, which would lead them to believe that p, whether or not p was actually true, this inevitability gives them reason to reduce confidence *only if* they also learn that:

- They would have based their belief on relevantly similar evidence in the counterfactual
- They would have based their belief on sufficiently good evidential processing abilities in the counterfactual

4.6 Inevitability and Evolution

We can now analyse inevitability in evolutionary debunking. I'll first outline the global, evolutionary debunking argument from inevitability, and then argue that their supporters are similarly unable to establish *Evidence Reveals Inevitability*.

Ruse and Wilson (1986, pp. 186–187) provide the earliest version of such an argument, contending that "even if external ethical premises did not exist, we would go on thinking about right and wrong in the way we do". Morton (2016, p. 242) argues similarly that "your belief that you have a categorical reason to (say) take care of your children would be adaptive regardless of whether you actually do have such a reason. The belief that there are any categorical reasons at all is similarly adaptive (and thus selected for)". Finally, Joyce's (2006, p. 181) belief pill case can also be cast in terms of inevitability, as detailed earlier.

These authors all hope to undermine the belief that categorical moral reasons exist, on grounds that this belief is evolutionarily inevitable. This inevitability can be represented as follows:

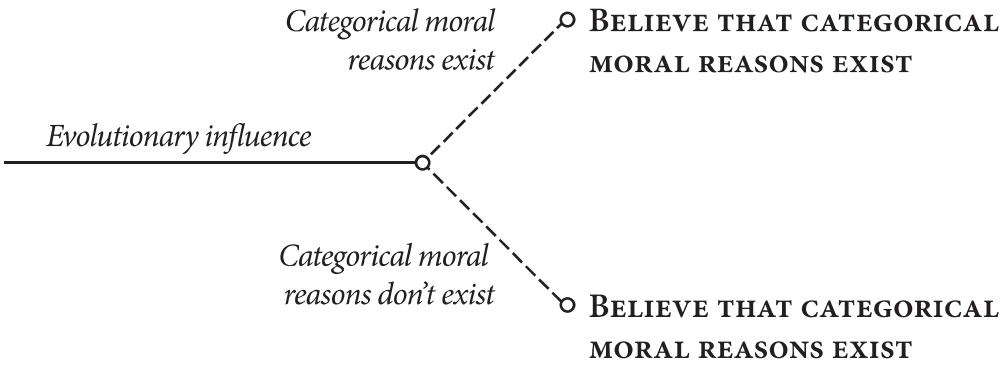


FIG 4.6 Inevitable belief that categorical moral reasons exist

Selective pressures would have ensured that we believe categorical moral reasons exist, regardless of whether such reasons actually do exist. The allegation is that upon learning this, you should reduce confidence in your belief in categorical moral reasons – just like how learning of the red light mandates reducing confidence in your red wall belief.

But does the evolutionary evidence show that our moral beliefs are inevitable in this way? I'll argue not. Think first about what counts as evidence for claiming that we would believe p regardless of whether p is true. First, we might observe that everyone in the actual world believes that p – this is a clue that there is a causal influence driving us towards the belief. But that influence might just be the very truth of p (White, 2010, p. 582), in which case the inevitability wouldn't be epistemically worrisome at all. To rule out this possibility, we should observe scenarios where the truth of p is *changed*, and check if we would still believe p then. If so, then we have a case of epistemically worrisome inevitability. For instance, in the red wall case, we might have done prior experiments shining the red light on red walls and white walls, and asked people about the colours of these different walls. If everyone believed that the wall was red, regardless of whether they saw a red wall or a white one, then we can be confident that the red wall belief is indeed rendered inevitable by the red light. Secondly, we might have a good understanding of the relationship between our beliefs and the relevant facts, which we could then use to predict that our belief that p would remain the same, even when the truth of p changes. Put in terms of the red wall case, we might have a deep understanding of how the human eye and brain work, and how red light interacts with the surface reflectance of different coloured walls. We could then infer that it will be difficult for

us to differentiate between a red wall and a white one when there's a red light shining on both of them.

Now return to debunking: do we have evidence for claiming that the belief in categorical moral reasons is evolutionarily inevitable? It doesn't seem so – neither of the above two lines of evidence for inevitability seem available to the debunker. First, it's not even true that everyone in the actual world believes that categorical moral reasons exist. The moral anti-realists are a clear counterexample: these anti-realists are humans too, and descended from the same ancestors who faced the same evolutionary forces that allegedly led to our tendency to believe in categorical moral reasons. And yet, they have managed to avoid the allegedly inevitable belief in categorical moral reasons.[39] This is analogous to discovering, in the red wall case, that even with the red light shining on the wall, some people nonetheless form the belief that the wall is white rather than red.

In response, the debunkers might shift their target, and say that what's inevitable is not the *belief* in categorical moral reasons, but rather the *intuition* that there are such reasons. The problem is, some anti-realists don't report having this kind of intuition either. Joyce (2009, p. 215) himself, in arguing against the intuitions of a moral realist, writes that:

> I'll grant arguendo the plausibility of the empirical thesis that most people share Huemer's clear and distinct realm of moral intuitions… But not everyone does. The moral nihilist, for example, does not. Conceivably, the moral nihilist has felt no tug in favor of morality at all (just as some atheists have never felt a glimmer of temptation towards religion).

But that's not all: we also need to observe the scenario where the truth of p is changed from what it is in the actual world, and ascertain that we still believe p then. In the moral case, however, it's hard to see how we could make such observations. Many philosophers think that

[39] Lahti (2003, pp. 647–648) uses the example of anti-realists in society to prove a related point: that belief in mind-independent moral facts is not necessary for facilitating cooperative behaviour.

if there are moral facts, these facts would be metaphysically necessary. So a world where there are no categorical moral reasons, but the non-moral facts remain the same, would be a world that is metaphysically impossible (but still conceptually possible). Our empirical evidence, however, comes only from the actual world – and it's unclear how such evidence could bear on what we would or wouldn't believe in the metaphysically impossible world. Moreover, we cannot intervene and change the moral facts in the way we might do with the colour of a wall. So it's difficult to see how we could obtain the first line of evidence – direct observational evidence – for the inevitability of our moral beliefs.

(My criticism here is related to, but different from, Clarke-Doane's (2016, pp. 26–28). He argues that our moral beliefs trivially satisfy the requirement of sensitivity – which roughly says that a belief is sensitive iff it's not the case that had the facts been changed, our moral beliefs would still remain the same. He thinks this because it's metaphysically impossible for the moral facts to be changed in the first place – so, trivially, it will be metaphysically impossible for such facts to be different, while our moral beliefs still remain the same. I don't go so far as to claim that the relevant requirement is trivially satisfied – I'm rather arguing that we *have insufficient empirical evidence* to prove such claims of inevitability or insensitivity.)

Secondly, as argued before, we don't have a good understanding of the evolutionary origins of our moral beliefs. So we also can't use this line of evidence to infer that we would have held a particular moral belief, regardless of what the moral facts are like – in the same way we might have done with the red wall belief and a deep understanding of human physiology and surface reflectance. In sum, it's not clear how the empirical evidence could show that we would have the same moral beliefs, even when the moral facts are changed – so we also don't have evidence of inevitability.

Moreover, the debunkers need to show that this inevitability is epistemically worrisome, by providing evidence that the relevant counterfactual selves would have used similar evidence, and sufficiently good evidential processing abilities, in forming their moral beliefs. But as with the case of chanciness, the current evolutionary evidence doesn't tell us much about

what these counterfactual selves would be like. Thus the two necessary conditions for inevitability to be epistemically worrisome are not satisfied either.

4.7 From Causal Inefficacy to Inevitability

Debates about inevitability centre around what we would believe when the moral facts are changed. This depends, in turn, on our models of what the moral facts are like. I believe that in pressing a challenge about inevitability, the debunkers assume a specific model of what the moral facts are like – namely, that such facts have no causal powers. If the moral facts have no causal powers, then the natural world – including what moral beliefs we would hold – would have gone on in exactly the same way, regardless of how the moral facts are like.

This conditional claim – that if the moral facts have no causal powers, then our moral beliefs are inevitable in an epistemically worrisome way – might well be correct. But I don't think the debunkers are licensed to assume the antecedent. After all, many moral philosophers do not accept this causal inefficacy. Some think that the moral facts *reduce to* some non-moral facts, some think that the moral facts *supervene on* non-moral facts, and still others think that the moral facts are non-natural but yet have causal powers (Shafer-Landau, 2012, pp. 25–29).

Moreover, even if the debunkers are licensed to assume causal inefficacy, the empirical evidence about evolution would be rendered redundant in the resulting debunking argument. First, this empirical evidence doesn't give us any further reason to think that the moral facts are causally inefficacious – the fact of causal inefficacy, if it is a fact, is not something that can be empirically established. Second, if we accept that the moral facts are causally inefficacious, then the inevitability of our moral beliefs would follow, regardless of what the causes of our moral beliefs are. The details of their causal origins – whether they were shaped by an innate mechanism, by social learning, or by inference – wouldn't matter, so long as the moral facts are causally isolated from the non-moral ones (Bedke, 2009; FitzPatrick, 2015, p. 892). At times, Joyce seems to concede this: in describing Harman's challenge (which he takes to be an inspiration for his own) he says that "[r]eference to biological natural selection is not

necessary to Harman's case; all his argument requires is that there is some complete explanation that can be given of our moral judgments for which their truth or falsity is irrelevant." (Joyce, 2006, p. 185)

Thus if the debunkers just assume a model on which the moral facts are causally inefficacious, they might get the verdict that our moral beliefs are inevitable – but this also renders the empirical evidence redundant. This redundancy of the empirical findings doesn't weaken this debunking argument (as I've noted in Chapter 3), but it does show that the argument doesn't add anything to old concerns about the causal inefficacy of the moral facts.

4.8 Conclusion

I've criticized *Evidence Reveals Flaw* as it manifests in global, evolutionary debunking arguments from chanciness and inevitability. Firstly, the evidence doesn't establish the chanciness of our belief in categorical moral reasons – we haven't observed suitably intelligent, social organisms who were subject to different evolutionary pressures and didn't believe in categorical moral reasons as a result. We also can't establish that this chanciness is epistemically worrisome, since we don't know much about what our counterfactual evolutionary selves would be like – what kind of evidence they would possess, or what their evidential processing abilities would be like.

Secondly, the evidence also doesn't show that our belief in categorical moral reasons is inevitable. This belief isn't even inevitable in our actual world – human anti-realists don't believe in categorical reasons, and don't have the intuition that such reasons exist, even though they are subject to the same evolutionary pressures as other humans. Moreover, our empirical evidence only pertains to the actual world, whereas establishing inevitability also requires finding out what beliefs we would have in the metaphysically impossible world where the moral facts are different from what they actually are.

The evolutionary origins of our moral beliefs are sometimes cast as either debunking (such that they show our moral beliefs to be in poor epistemic standing) or vindicating (such that they show them to be in good epistemic standing). There is, however, a third position that deserves serious consideration – it holds that the current evidence about evolutionary origins just *doesn't support any clear epistemic verdict*. The evolutionary evidence is not ready-made for drawing epistemic conclusions, and we shouldn't pretend otherwise.[40]

[40] But see Isserow (2018), who argues that the current evidence still suffices for denying moral knowledge.

5 Probability Weighting in Ethics: A Case Study for Debunking

You are operating a ship alone in a furious storm. You receive a call for help from a remote island, where you find 60 people awaiting rescue. However, your ship is certified to carry less than 60, so you are faced with two options: attempting to rescue all 60, which has a 90% probability of succeeding, but a 10% probability that the rescue will fail and all 60 will die; or, you can attempt to rescue exactly 54 people, which has a 100% probability of success, such that 54 people will live but 6 will die. Which option should you go for? Intuitively, it seems like you should go for the safe option of saving the 54 for sure. Intuitions of this kind – concerning the provision of aid in risky situations – are invoked in the philosophical literature. But are these intuitions trustworthy? In this chapter, I hope to make progress on this question, by using intuitions about risky aid as a case study for debunking.

Why care about these intuitions? Because most, if not all, attempts to aid others will have some risk of failing. Any moral theory must give guidance about how to evaluate the risks associated with different courses of action – that is, any moral theory must tell us which risk attitude we should adopt in our moral choices. In the course of theorising about such risk attitudes, we will likely have to appeal to – or undermine – various intuitions about risky aid. So whatever moral theory you subscribe to, you should be concerned with whether such intuitions can be debunked[1] (and are hence untrustworthy), or whether they are more likely accurate reflections of the moral facts.

This chapter is organized as follows: I begin by clarifying these cases of risky aid. I then outline some relevant empirical research about our preferences over different monetary

[1] In this chapter, I speak of intuitions (rather than beliefs) being debunked. This is because there isn't much literature about which risk attitude we should adopt in morality, hence it seems more natural to talk of intuitions rather than settled belief. In any case, any mention of intuition in this chapter can be substituted with "belief based on intuition" without any substantive difference.

gambles, and argue that our moral intuitions about risky aid seem to follow a pattern that's similar to what's been observed with these preferences. I then explore two possible normative implications that might be supported by this pattern: first, it might support debunking arguments which aim to undermine our intuitions about risky aid – I explore a few such arguments, and consider how the regress and redundancy objections from Chapter 3 might manifest here. Second, this pattern might inform a positive, vindicatory argument which assumes that our moral intuitions are generally reliable. I lean toward the debunking conclusion rather than the vindicatory one – but I also believe that whichever normative implication you wish to draw, the empirical findings will usefully inform and supplement your moral theorising.

5.1 Risks and Rescues

Think again about the case which I began with – which I'll call *Large Ship*.

> Large Ship
>
> Risky Rescue: 90% probability of leading to 60 people surviving (and none dying); and 10% probability of no one surviving (and all 60 dying).
>
> Safe Rescue: 100% probability of leading to 54 people surviving (and 6 dying).[2]

We can calculate the expected value of each rescue attempt by weighting the value of each possible outcome by the probability that that outcome occurs. If we assume that each life rescued has the same moral value (an assumption I'll relax later on), then the two options have the same expected value:[3]

[2] Assume that there are 54 *specific* people who will live if we attempt this safe rescue, and 6 specific people who will die. Imagine, perhaps, that the 54 are in one container, and 6 are in another, and as the captain you can take one or both containers on your ship.

[3] In general, I focus on options with the same expected value, in order to isolate the effects of risk – separate from the effects of expected value – on permissibility. Also, I'm happy for this expected value to be calculated after we've ruled out possibly irrelevant (risks of) harms, like risks of headaches when compared against deaths

EV(Risky Rescue) = 90%*U(60 alive, 0 dead) + 10%*U(0 alive, 60 dead) = U(54 alive, 6 dead)

EV(Safe Rescue) = 100%*U(54 alive, 6 dead) = U(54 alive, 6 dead)

Nonetheless, we seem to have the intuition that we're morally required to go for Safe Rescue, and are morally forbidden from trying Risky Rescue.[4] The Risky Rescue seems like too much of a gamble, since the extra chance of saving everyone is not enough to justify choosing Risky Rescue over Safe Rescue. Even when these two options have the same expected value, it seems like the success probabilities of these rescues matter too, morally speaking. In what follows, I will further explore how our intuitions about such cases change in response to the success probabilities.

But first, some clarifications about cases like these. First, I'm concerned with distinctively *moral* intuitions about what we should do in situations where our rescue attempts have different probabilities of success. This is different from Buchak (2017), who uses *prudential* intuitions about how we should rationally choose in different gambles, along with normative premises about the Rawlsian veil of ignorance, in order to derive a conclusion about distributing goods according to relative prioritarianism. Moreover, I'm concerned only with the *moral permissibility* of different actions, rather than with blameworthiness for performing an action, or with the moral goodness of the action itself. Second, the probabilities I'm concerned with are *justified degrees of belief* (or evidential probabilities) that a moral agent has about the situation – these degrees of belief are supported by the evidence, or are otherwise epistemically rational to hold. An agent's justified degrees of belief will most naturally affect the subjective permissibility of their action (permissibility in light of the agent's beliefs/evidence), but they might also affect its objective permissibility (permissibility

(Lazar, 2018). In the cases that I consider, relevance is not an issue because the harms at stake are all of the same order (they're all deaths).

[4] Even if you didn't have this intuition, this chapter could still contribute to your understanding of how to debunk such intuitions (when others report them, for instance).

in light of all the facts) (Lazar, 2019, pp. 5–10).[5] Third, these justified degrees of belief are about the success probabilities of different courses of action – how likely each rescue attempt will succeed at saving the people – rather than other probabilities, like the probability that the people would have ended up in specific groups or locations in the first place.[6] Fourth, I'm concerned only with choosing between two different actions, and will set aside the question of what to do if we can instead randomize between the two actions – for instance, by rolling dice to decide.[7] Fifth, as is customary when discussing these cases, I'll assume that there are no further reasons distinguishing the people who might be saved by the different options – we don't have special obligations to any specific group, no specific group is responsible for this predicament, and so on.

Finally, I'm using the somewhat awkward locution of 'x alive, y dead' to characterise the possible outcomes, because I want to avoid unwanted implicatures or framing effects in the specification of my cases. Merely saying that "20 people are alive", for instance, while leaving out the number who are dead, leaves out important information that might influence our evaluation of the case (Kühberger, 1995). It might even imply that *at least* 20 people are alive, (Mandel, 2014), which I want to avoid.

5.2 Risk Aversion and Risk Seeking in Morality

With these clarifications in place, we can return to think about our initial Large Ship case. We can distinguish between three ways in which the risks involved in different options affects the permissibility of those options, when such options have the same expected value:

---

[5] To be precise, these are justified degrees of belief about the likelihood of each possible outcome, and not about the likelihood that a specific person will be harmed/saved (which Otsuka (2015, pp. 80–81) argues – rightly, I believe – has no moral relevance).

[6] Here I follow Rasmussen's (2012) terminology. She has a sophisticated argument for why when we also factor in the probability that each person would have ended up in specific locations/groups, we get an approach that's extensionally equivalent to maximising expected value. My cases, however, bear more on the question of whether there's a tiebreaker for two options that have the same expected value, which is a question she leaves open.

[7] There is much work on this: see, for instance, Taurek (1977), Kamm (1993), Rasmussen (2012), Walden (2014).

Moral risk aversion: When two options to aid have the same expected value, their relative riskiness creates a moral reason in favour of the *less risky* option (if there is one).

Moral risk neutrality: When two options to aid have the same expected value, their relative riskiness does not create any reason in favour of either option.

Moral risk seeking: When two options to aid have the same expected value, their relative riskiness creates a moral reason in favour of the *more risky* option (if there is one).

The intuition that we should go for Safe Rescue over Risky Rescue appears to support moral risk aversion, since Safe Rescue is the less risky of the two options. We could argue for this moral risk aversion in the following ways: we might think that when we don't know who the people on the island are and what their risk preferences are like, we should err on the side of caution and choose the safer option to aid;[8] or we might argue that we have a duty to guard against the worst possible outcome, where we don't save anyone and all 60 die.[9]

So just reflecting on our intuition about Large Ship, we might conclude that when it comes to aiding, morality is risk averse. However, the relevant empirical literature suggests additional cases that might be worth considering, and proposes an alternative explanation of our intuitions that is worth taking seriously.

[8] This draws on Buchak (2017, pp. 21–24), although she is concerned with prudential (rather than moral) justification, and with choosing between different options that only affect a single person (rather than options that affect different people).

[9] Rawls (1999, pp. 130–139) argues for his maximin principle in quite a different context. Altham (1983, p. 21) points out, however, that Rawls's defence of maximin relies on very specific conditions – for instance, the condition that the probabilities of different outcomes cannot be computed – so Rawls' arguments might not be useful for thinking about the ethics of risk more generally.

5.2.1 The Empirical Findings

Behavioural economists have studied peoples' preferences with regards to different monetary gambles, in the hopes of obtaining a descriptively accurate picture of how we make decisions about risky prospects. These economists conduct experiments, like offering subjects different gambles – for instance, a gamble that has a 5% probability of paying $100 – and asking them how much they would be willing to pay for that gamble (Gonzalez & Wu, 1999), or by giving subjects a choice between two gambles and asking them which they would prefer (Fehr-Duda & Epper, 2011). Outside of monetary gambles, economists have also studied our preferences with regards to risky policy decisions. For instance, Tversky and Kahneman (1981) use the famous Asian Disease problem to study the effect of framing policy options in terms of gains (number of lives saved) versus in terms of losses (number of people dying).[10]

These economists argue that the evidence supports what's known as Prospect Theory – an alternative to expected utility theory that purports to describe human decision-making more accurately. Prospect Theory postulates a value function that distinguishes between losses and gains about a reference point, and a probability weighting function that transforms the stated probabilities into decision weights (Barberis, 2013). Much has been made about how Prospect Theory's value function provides an alternative, and potentially debunking, explanation for our moral intuitions about doing and allowing (Horowitz, 1998; Van Roojen, 1999; Kamm, 1998; Sinnott-Armstrong, 2007; Dreisbach & Guevara, 2019). But I'll set that debate aside, by considering only options that have value in the gains domain, which could be used to model our duties to aid.[11] I will instead focus on Prospect Theory's probability weighting function, which purports to describe how people process the probabilities stated by the decision

[10] There are two key differences between the classic Asian Disease Problem (ADP) cases and my own. First, I specify both the number of people alive and the number dead – whereas risky choices in the ADP only talk about the number saved of lives saved, or the number of lives lost, but not both. This makes the ADP subject to problematic implicatures detailed earlier. Second, the risky choices in the ADP uses outcome probabilities of 33.3%, which minimizes the probability weighting effect that I wish to study – in contrast I look at more extreme probabilities where this effect will likely be present.

[11] Dreisbach and Guevara (2019, pp. 621–623), Spector (2016, p. 124), and Daniels (2015, p. 120) make comments suggesting this.

problem. The theory claims that we overweight low probabilities (roughly, probabilities lower than 33.3%), and underweight moderate to high probabilities, in comparison to expected utility theory.[12] (Probabilities of 0% and 100% are unaffected by weighting.) This pattern is not due to erroneous beliefs – people understand what the stated probabilities are – but are rather due to decision weights that people attach to different outcomes (Barberis, 2013, pp. 177–178; Tversky & Kahneman, 1992).This weighting can be summarised in the graph below, where the horizontal axis denotes the probabilities stated in a decision problem, and the vertical axis denotes the assigned decision weights. The dotted 45-degree line shows how expected utility theory weighs probabilities linearly – if a decision problem specifies a probability to be x, the decision weight assigned will also be x. The solid line shows how Prospect Theory weighs the stated probabilities:

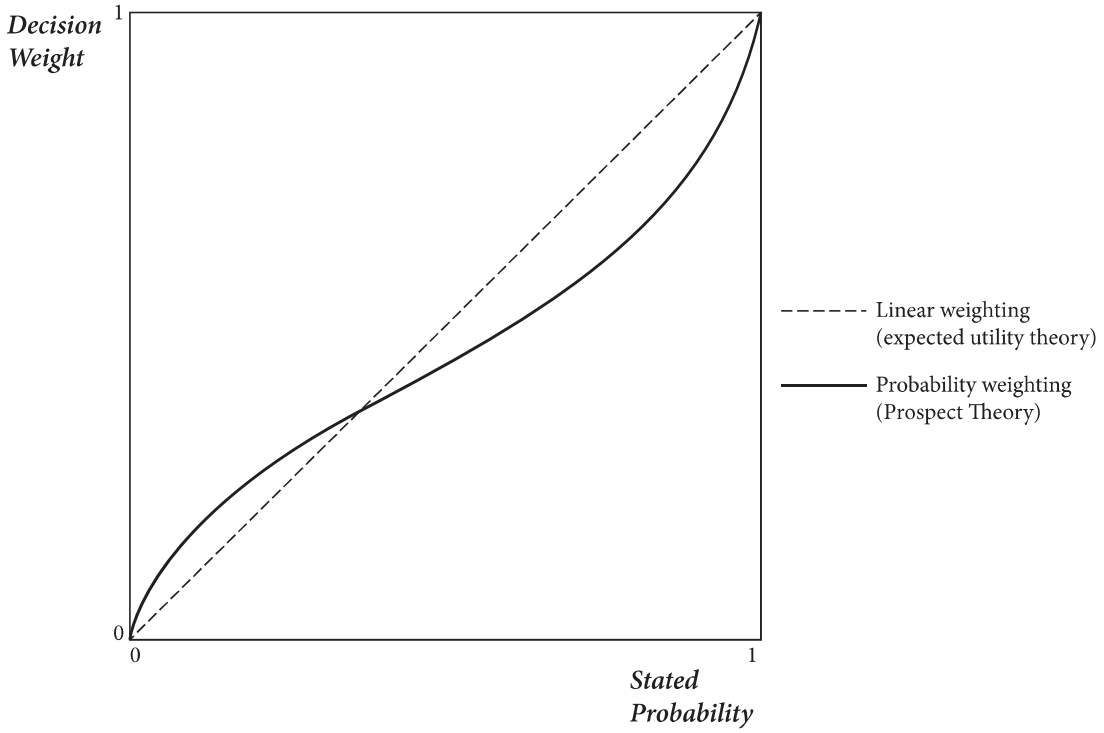


FIG 5.1 LINEAR WEIGHTING VS PROBABILITY WEIGHTING

[12] To be clear, I'm using Cumulative Prospect Theory's S-shaped weighting function, but to transform outcome probabilities rather than cumulative probabilities. In Cumulative Prospect Theory, the probabilities that are transformed are not the probabilities of each outcome, but rather the cumulative probability – of getting that outcome or better (for gains), or of getting that outcome or worse (for losses). (See also Buchak (2013), who uses the cumulative approach in a theory of prudential rationality.) Using cumulative probabilities allows us to respect stochastic dominance in many-option cases, and is perhaps more empirically adequate (Fennema & Wakker, 1997). But we are only dealing with two-option cases, so we can just work with the non-cumulative version of Prospect Theory, which I believe will work just as well.

As you can see, in contrast to expected utility theory's linear weighting, Prospect Theory probability weighting function postulates that we overweight low probabilities and underweight moderate to high ones. This weighting function is used to explain the following pattern of preferences: subjects were on average willing to pay $63 for a gamble that had 90% probability of paying $100 and 10% probability of paying $0 – this is risk aversion, since it implies that subjects prefer getting the gamble's expected value ($90) for sure to getting the gamble itself. On the other hand, subjects were also on average willing to pay $10 for a gamble that had 5% probability of paying $100 and 95% probability of paying $0 – this is risk seeking, since it implies that subjects prefer the gamble itself to the sure guarantee of obtaining its expected value (a mere $5) (Barberis, 2013; Gonzalez & Wu, 1999).[13] Part of the economists' explanation for this is that in the first gamble – with a 90% probability of paying $100 – we underweight the 90% probability involved, while in the second gamble – the one with a 5% probability of paying $100 – we overweight the 5% probability. This theory of probability weighting was formulated to explain preferences between different monetary gambles, but could it also explain our moral intuitions about risky aiding? I believe it might. In Large Ship, we were choosing between a Risky Rescue that had 90% probability of rescuing 60 people and 10% probability of rescuing none, versus a Safe Rescue with the 100% probability of rescuing exactly 54 but also leading to 6 deaths. We might also have underweighted the 90% probability of success in the Risky Rescue – just like the experimental subjects did with the monetary gamble that had a 90% probability of paying $100 – and this led to moral risk aversion.

5.2.2 An Example of Moral Risk Seeking

If Prospect Theory's probability weighting function correctly explains our intuitions about risky aiding, then we should be able to find a case of moral risk *seeking* too – a case that's

[13] To be clear, this pattern – risk seeking at low probabilities, and risk aversion at high probabilities – happens with gains, and is actually reversed for losses. See, for instance, Kahneman (2013, p. 317) on the fourfold pattern of risk.

analogous to subjects' willingness to pay $10 for a gamble that only has a 5% chance of paying $100.

There do seem to be such cases. Suppose, for example, we modified our initial case so that now, your ship is much smaller – such that you can only rescue exactly 6 people with 100% probability, or, alternatively, you can try to rescue all 60 people with a mere 10% probability of success. The options in this new case – which I'll call *Small Ship* – are as follows:

> Small Ship
>
> Risky Rescue: 10% probability of leading to 60 people surviving (and none dying); and 90% probability of no one surviving (and all 60 dying).
>
> Safe Rescue: 100% probability of leading to 6 people surviving (and 54 dying).

Again, if we assume that there is no diminishing marginal moral value of lives, these two options have the same expected value:

> EV(Risky Rescue – Small Ship) = 10% *U(60 alive, 0 dead) + 90% *U(0 alive, 60 dead) = U(6 alive, 54 dead)
>
> EV(Safe Rescue – Small Ship) = 100% * U(6 alive, 54 dead) = U(6 alive, 54 dead)

However, it seems here that Risky Rescue is morally required – or at least, morally permissible. While Safe Rescue guarantees that exactly 6 people will be rescued, the chance of rescuing all 60 people does seem to provide a strong reason in favour of the Risky Rescue. This is moral risk seeking (or at least, moral risk neutrality[14]) – since we have two options with the same expected value, but the risks generate a reason supporting the *riskier* option.

[14] If we thought that Risky Rescue was merely permitted but not required in Small Ship, then this will just be a case of moral risk neutrality. I will work with the stronger version of moral risk seeking here, but my arguments will go through equally well if our intuition in Small Ship just exhibited risk neutrality.

Different theoretical justifications have been offered for similar intuitions: Kamm (1993, pp. 124–125) entertains (but ultimately rejects) the idea that people have a right to a non-zero probability of survival, while Dreisbach and Guevara (2019, p. 619) argue that it might be good to give everyone a reasonable probability of survival. Kamm (1993, pp. 124–126) argues that when it is possible to save everyone with some significant and proportionate[15] probability, we can try that over saving some subset with certainty – because it gives each person a chance according to their proportionate moral weight, and is a form of solidarity with others at risk. Daniels (2015, p. 118) invokes a related case[16] to argue that we should disperse risks of harm over a larger group of people, rather than concentrate them on a few people – the same might apply for the probability of survival. And Dreisbach and Guevera (2019, p. 620) argue that other things being equal, it's better to let chance decide – this can be understood as claiming that it's better for people to have a probability of survival that's neither 0% nor 100%.

The intuition in Small Ship is that we should try the risky rescue – this is a risk seeking choice. But now compare that with our initial Large Ship case – where we had to choose between a risky rescue with 90% probability of rescuing all 60, and a safe rescue with 100% probability of rescuing exactly 54. The Large Ship case exhibited the exact opposite kind of verdict – moral risk aversion – since our intuition there was that we're required to go with the safe rescue, and we're forbidden from going for the risky one. This pattern – where we flip from risk seeking to risk aversion as the probability of gains increases past an inflection point

[15] Proportionate probabilities are ones that are proportionate to the moral weight of each group that could be potentially saved. Kamm (1993, pp. 123–124) allows, for instance, that if we can save an additional person by lowering the survival probabilities of an initial group, we can do so in proportion to the moral weight of that additional person. For instance, if we could save an additional person by lowering the probability of survival of five who would have been saved with certainty, we can lower their probability by 1/6 – but the probability of survival of the initial group should never go below the probability of survival of the additional people. Extrapolating to the case of Small Ship, I believe each new person added as we move from the 6 (who would have survived for sure under the Safe Rescue) to all 60 will permissibly bring the probability of survival of everyone down to 10%.

[16] His case involves choosing between the sure chance that one person will die, versus one of five people possibly dying, giving a 20% probability of death to each – but it's *certain that exactly one person will die* under this risky option. In my cases, however, there is no similar certainty of death – it is possible that the risky rescues in Small Ship and Large Ship succeed, which would lead to no one dying.

(while the expected value remains the same) – is exactly what's predicted by Prospect Theory's probability weighting function. So the theory does seem descriptively accurate, at least when applied to the intuitions about these two cases. But this is a mere descriptive claim about how our intuitions change in response to different success probabilities. Can we conclude anything further about whether these intuitions are trustworthy or not? I turn to this issue in the next section.[17]

5.3 The Epistemic Status of Our Intuitions about Risky Aiding

5.3.1 Debunking Argument from Linear Weighting

Given the pattern observed, we might offer the following debunking argument against our intuitions about risky aid. This argument assumes that morality weighs probabilities in a linear fashion – for instance, the increase in the strength of moral reasons for an option when we move from 10% probability of rescue to 20% probability of rescue should be the same as the increase when we move the probability of rescue from 50% to 60%.[18] But it observes that our intuitions about risky aiding don't seem to follow this linear pattern of weighting – because low probabilities are overweighted, and moderate to high probabilities are underweighted. Therefore, this argument concludes, our intuitions about risky aid are untrustworthy.

While there might be a plausible case to be made that morality weighs the probabilities linearly,[19] this debunking argument is vulnerable to the argument redundancy objection explored in Chapter 3. This debunking argument already assumes that morality weighs

[17] Strictly speaking, the vindicatory and debunking arguments of the next section don't even require that Prospect Theory be true in its entirety. These arguments only need us to have the pair of intuitions concerning Large Ship and Small Ship. Thanks here to James Willoughby.

[18] Rasmussen (2012, pp. 210–211) entertains a similar argument in a different context, but she compares the move from *zero probability* to a non-zero probability, with the same-sized move from a non-zero probability to a higher non-zero probability. Also see Oddie and Milne (1991) and Lazar (forthcoming).

[19] For instance, we might argue that any nonlinear weighting of probabilities will commit us to the moral version of a Dutch Book argument (for example, see Quiggin's (1993, pp. 121–124) explanation of how nonlinear weighting might lead us to dynamic inconsistency). Thanks here to James Willoughby.

probabilities linearly, and all its evidential impact comes from this assumption. The undermining of our intuitions about risky aiding, even if it's achieved by this debunking argument, doesn't create any *further* evidential impact. We can also think of the problem in terms of begging the question. A philosopher who uses this debunking argument to undermine another person's intuitions seems to beg the question against that person, since their argument just assumes that this person's intuitions were unreliable to begin with.[20]

This same problem of evidential redundancy plagues debunking arguments which assume from the outset that morality is risk averse (or that it is risk seeking). For instance, we might follow Buchak (2017) in arguing that when we don't know what the risk preferences of others are, we should err on the side of caution and choose the less risky option. We could then conclude that the *risk seeking* intuition in Small Ship is debunked. But if we have already admitted some independent basis for claiming that morality is risk averse, this same basis doesn't create further evidential impact by also undermining our risk seeking intuitions.

5.3.2 Debunking Argument from Constant Contribution

The earlier debunking argument from linear weighting wasn't evidentially impactful, because its assumptions were too strong – such that while it might undermine our intuitions, it only does so in a context where such undermining is evidentially redundant. If we are to debunk our intuitions about risky aiding in an evidentially impactful manner, we need a weaker assumption that could still generate the debunking conclusion. One such assumption might be that the risks involved can only generate reasons supporting options of at most one kind. In particular, the risks can generate a reason supporting the less risky option, or a reason supporting the riskier option, or a reason supporting neither – but these risks cannot support more than one kind of option. In our intuitions about risky aiding, however, the risks seem to generate reasons for two kinds of options – at low probabilities (as with Small Ship), we appear to have a reason favouring the riskier option, but at moderate/high probabilities (as in

[20] While I have some sympathy for this way of putting the point, I prefer the earlier version – which works in terms of evidential redundancy – since that version doesn't involve an interlocutor against whom we beg the question, and so fits better with the personal view of debunking introduced in Chapter 2.

Large Ship), we appear to have a reason favouring the less risky option. So, our moral intuitions about risky aiding are untrustworthy.

This variable contribution of risk can be made vivid by applying the theoretical justifications offered for one type of case (say, Small Ship) to generate a verdict in the other case. Recall that in Small Ship, we had the intuition that the risky rescue with 10% probability of rescuing all 60 can be permissibly chosen over the 100% probability of rescuing 6. Kamm could justify this by saying that the risky rescue gives everyone a significant and proportionate probability of survival, Daniels could argue that this risky rescue disperses the probability of survival across a larger population than the safe one does, Dreisbach and Guevara could argue that the risky rescue lets chance decide the fates of all 60 people, as opposed to the safe rescue which lets chance decide for no one. But if we apply these same justifications to *Large Ship* instead – where we had to choose between the risky rescue with a 90% probability of rescuing all 60, and the safe rescue with a 100% probability of rescuing exactly 54 – we get the verdict that we are permitted to go with the risky rescue too, which is counterintuitive. It's worth going through each justification again: the risky rescue in Large Ship also gives everyone a significant and proportionate probability of survival (at 90%!); it disperses the probability of survival to 60 people (rather than concentrating it on 54 like the safe rescue does); it lets chance decide the fates of 60 people (as opposed to the safe rescue, which lets chance decide for no one). And yet, we don't have the intuition that Risky Rescue is permissible in Large Ship.[21] Thus there seems to be a deep tension between our intuitions in Small Ship and Large Ship, if we think that risk can only contribute a reason for one kind of risky option. And, to that extent, such intuitions can be debunked. This debunking argument can be cast as one from chanciness – because our moral intuitions change even when the moral facts (about which constant risk attitude is required by morality) remain the same. More formally, the argument can be phrased as below:

[21] We could do the same by applying the justifications for moral risk aversion to Small Ship, but I'll leave that out for brevity.

(Constant Contribution Assumption) Either morality is risk averse, or it is risk seeking, or it is risk neutral throughout different levels of success probabilities.

(P1) If morality is risk averse, then the Small Ship intuition is mistaken.

(P2) If morality is risk seeking, then the Large Ship intuition is mistaken.

(P3) If morality is risk neutral, then both the Small Ship and Large Ship intuitions are mistaken.[22]

(C1) Therefore, at least one of the Small Ship and Large Ship intuitions is mistaken. (from the Constant Contribution Assumption, P1, P2, P3)

With the Constant Contribution Assumption, we narrow down the possibilities to three – the morality of aid is either risk averse, risk seeking, or risk neutral across different success probabilities. And with each of P1-P3, we show that our intuitions would be mistaken in each of these possibilities. For instance, P1 states that if morality is risk averse, then one of our intuitions (about Small Ship) is mistaken because it supports moral risk seeking instead. We do the same with P2 and P3, to show that regardless of whether morality is risk averse, risk seeking, or risk neutral, at least one of these intuitions is mistaken. So while we might not know which of the three disjuncts of the Constant Contribution Assumption is true, we can still get the debunking conclusion. Notice that the Constant Contribution Assumption is a much weaker assumption than the previous assumption that morality weighs probabilities linearly. This allows for a debunking argument that isn't evidentially redundant (or question-begging) in the way that the earlier argument from linear weighting was.

[22] Here I'm assuming the stronger intuition for Small Ship – that we're morally required to go for the risky rescue. If we only thought it was morally permissible, then only the Large Ship intuition runs afoul of moral risk neutrality.

5.3.3 Debunking Argument from Implausible Change in Contribution

The Constant Contribution Assumption is a weak assumption that generates a plausible debunking argument. But I believe that we can reach the debunking conclusion with an even weaker assumption. Here's how: we could allow that the risks make variable contributions at different levels of probability – such that it's possible for morality to flip from risk seeking to risk aversion – but still argue that the way this contribution varies is implausible.

Consider the only kind of justification offered so far that has a threshold concept which can be used to accommodate the flip from risk aversion to risk seeking – Dreisbach and Guevara argue that it might be good to give people what they call a respectable (as opposed to merely nonzero) chance of survival, and Kamm argues that we can give everyone a proportionate and significant probability of survival. They might argue then as follows: low probabilities of survival don't qualify as respectable or significant, and only moderate to high probabilities of survival do. The problem, however, is that this justification predicts a flip in the opposite direction from what we actually observe: it predicts that at low probabilities, as in Small Ship, we would be morally risk averse – since the 10% probability of rescuing everyone in the risky rescue wouldn't count as reasonable or significant. And at high probabilities, as in Large Ship, we would be morally risk seeking – since the 90% probability of the risky rescue's saving all 60 would count as reasonable and significant. But our intuitions went the other way instead – in Small Ship, we were risk seeking, and in Large Ship, we were risk averse. So even if we allowed that the contribution of risk could change at different levels of probability, the observed direction of change might still be morally implausible, and thus provide grounds for a debunking argument.

Moreover, even if there is a plausible moral explanation for why the risk attitude changes at different levels of probability, this explanation must justify the inflection point of this change. At present, the empirical results suggest that the risk attitudes change at the inflection point of 33.3% – but it is unclear why this specific probability would be morally significant.

5.3.4 Diminishing Marginal Moral Value of Lives

The same point – about the wrong direction of influence – also applies to the putative explanation that there is diminishing marginal moral value of lives saved. This explanation says that the moral value of an additional life saved decreases, as the stock of lives already saved increases. For instance, the value of the additional life saved when we move from 53 to 54 lives saved is more than the value of the additional life when we move from 59 to 60 lives saved. This could be used to explain the Large Ship intuition in isolation: the additional lives saved when the Risky Rescue saves everyone (60 people) are worth proportionately less because of diminishing value, so we should go with the Safe Rescue in Large Ship, which saves 54 for sure. However, this explanation cannot *also* account for the moral risk seeking exhibited in Small Ship – in fact, the explanation implies that we have more reason to choose the Safe Rescue in Small Ship than we do in Large Ship. This is because the Safe Rescue in Small Ship only saves 6 lives for sure – and there will be much less diminishing value going on when just 6 lives are saved, as compared to when 54 lives are saved (in the Safe Rescue of Large Ship). (Diminishing marginal moral value doesn't affect the relative values of the risky rescues of these two cases, because the risky rescues have the same possible outcomes.) Hence diminishing marginal moral value cannot account for *both* the risk aversion in Large Ship and the risk seeking in Small Ship.

5.3.5 Issues and Implications in Debunking

I'll now highlight some further features of the debunking arguments just considered, and point out some potential issues. First, these debunking arguments are quite different from the evolutionary debunking arguments considered in Chapter 4. For one thing, they are local rather than global – these arguments only purport to undermine intuitions of a specific kind, not all our moral intuitions or beliefs at once. For another, the debunking arguments here draw on experimental evidence from behavioural economics, rather than evidence from evolutionary psychology. As I've argued in Chapter 4, the experimental evidence better

establishes the problematic chanciness of the target moral beliefs, and is more likely to support a debunking argument

Secondly, these debunking arguments avoid the findings redundancy objection presented in Chapter 3 – which, recall, alleged that the empirical findings are redundant in debunking arguments, and that such arguments are really just armchair reasoning in disguise. Here, the findings from behavioural economics have been crucial in uncovering the variable contribution of risk at different probabilities. Notice, for instance, that while moral philosophers have detected either moral risk-seeking or risk aversion in our intuitions about risky aiding, none have so far proposed that the contribution of risk flips when we consider low versus moderate/high probabilities of aid. The discovery of this novel pattern then provided the basis for debunking. Of course, we could in principle have detected this pattern with careful consideration of the cases alone – as we've done with Large Ship and Small Ship. So, in a sense, the empirical findings are redundant – because we *could have* reached the same conclusion from the armchair. Nonetheless, we *didn't actually* do so – the empirical findings were still crucial for discovering this pattern in the first place, hence such findings have still played an important role in our moral theorising.

Thirdly, some might worry that the debunking arguments presented here are subject to the regress objection, also considered in Chapter 3. For instance, the Constant Contribution Assumption might be too similar to our intuitions about the Large Ship and Small Ship cases, so this assumption might itself be debunked too. I can't fully defend against this objection here, but I can offer two quick responses: firstly, the Constant Contribution Assumption is a more abstract principle – it concerns how the risks contribute to the permissibility of options – and is at a different level of generality from our Large Ship and Small Ship intuitions, which are intuitions about what we should do in specific cases. This might create a sufficient difference between the two to stop the regress going, if we thought that a intuition/belief's epistemic status is correlated with its level of generality, as Huemer (2008) does. Secondly, the Constant Contribution Assumption seems so far to be assumed by everyone in the literature,

whether they endorse risk aversion, risk seeking, or risk neutrality.[23] This highlights how uncontroversial the assumption is – although of course, that doesn't entail that the assumption is right.

Now turn to the implications of debunking: suppose we've undermined our intuitions about Large Ship and Small Ship. What now? First, there's a question of scope: does it follow then that *all* our moral intuitions concerning risk are undermined – regardless of whether they're about aiding or about harming? Or is it just our intuitions about risky aiding specifically? Or, more narrowly still, our intuitions about risky aid when the expected value of two options is the same? These are difficult questions that I cannot settle fully here. We need to do more empirical work to see if intuitions about other cases are subject to the same flipping pattern – for instance, in cases where the expected value of two options to aid are different (these cases are especially difficult because we need to tease apart the influence of the expected value from the influence of probability weighting). If we still observe this characteristic flipping between risk seeking and risk aversion at different levels of probability, then intuitions of the broader type are debunked as well. I'm inclined to adopt a conservative attitude and consider only the most fine-grained class of intuitions debunked for now. That is, I am inclined to only debunk intuitions about two risky aiding options that have the same expected value, and wait for further empirical results about broader classes of intuitions.

Secondly, there's a question about correction.[24] Suppose we are quite confident in the Constant Contribution Assumption, and have determined that at least one of the Large Ship and Small Ship intuitions are inaccurate. Could we further improve our epistemic position here? If we were more confident in our intuition about one of these cases than the other, then we can combine that with the Constant Contribution Assumption, and generalise the risk attitude to the case that we're less confident about. For instance, I am more confident in my intuition about Large Ship than about Small Ship, so I might disregard the Small Ship intuition and conclude that morality is risk averse throughout. Alternatively, if we were

[23] Thanks here to James Willoughby.

[24] Vavova (forthcoming) raises the possibility of correction in the context of evolutionary debunking.

equally confident in both these intuitions, then we might have to suspend judgment for now, and look for other theoretical justifications – independent of our intuitions – for why we should adopt one risk attitude over another in cases of risky aiding.

5.3.6 Positive Argument from Empirical Findings

Alternatively, you might find the Constant Contribution Assumption implausible, and instead be much more confident in your intuitions about both Large Ship and Small Ship. In this case, the empirical findings might support a vindicatory argument of the following form: our moral intuitions are reliable indications of the moral facts, and our intuitions follow pattern X. Therefore, the moral facts follow pattern X (Kahane, 2013, 2016, pp. 300–301). Applied to the case of probability weighting, we might use the empirical results to conclude that morality also weighs probabilities in the distinctive manner postulated by Prospect Theory – overweighting low success probabilities of aid, and underweighting moderate to high ones. While this might not sound like an exciting conclusion, it still reveals how the empirical evidence could usefully supplement ethical theorising. We might rely on the sophisticated experimental methods and sound empirical reasoning of the social scientist in order to detect subtle, nonobvious patterns in our moral intuitions, and these novel patterns then inform our theories about what the moral facts are like (Kahane, 2013).

5.4 Conclusion

In this chapter, I've examined a new case study for debunking, concerning our intuitions about risky aiding. Using empirical evidence and theoretical explanations from behavioural economics, I devised two cases – Large Ship and Small Ship – which showed how we seem to flip from risk aversion to risk seeking when we move from low to moderate/high probabilities of aid (while the expected value of the rescue remains the same).

I then looked at possible normative implications of this descriptive picture. First, we could try a debunking argument which assumes that morality weighs probabilities linearly. I've argued, however, that this debunking argument is evidentially redundant, since all its evidential impact comes from its initial assumption about linear weighting, rather than from its undermining of the relevant intuitions. Second, we could try a debunking argument from the constant contribution of risk: we might argue that the risks can only generate reasons supporting at most one kind of option: supporting the riskier option in a choice set, supporting the less risky option, or supporting neither. Third, I looked at further debunking arguments which allow risk attitudes to change across different levels of probability, but still contend that the direction of change observed is morally implausible. I argued that these latter two arguments are more convincing, because they don't assume the falsity of the target intuitions. I then also looked at further issues with debunking – to do with the empirical evidence used, the threat of regress, the scope of the debunking, and the potential for correction. If you thought that the assumptions of these debunking arguments were implausible, however, then you might use the empirical findings to support a vindicatory argument instead: our moral intuitions are generally reliable, and they follow this pattern of probability weighting, therefore morality itself weighs probabilities in this distinctive manner. Whether we take the debunking or the vindicatory route, the empirical findings have helped by uncovering new cases and proposing new theories about what our intuitions might be responding to.

In recognising some of these points at a more general level, Kahane (2013, p. 422) concludes "empirical research will take over part of ethics. The armchair will be pushed to the corner." While I agree with him in substance, I prefer a different metaphor – we are instead upholstering the armchair to keep up with the times. For it is still the moral philosopher, sitting in their improved armchair, who must decide whether to keep or to debunk the relevant intuitions.

6 A Bayesian Analysis of Debunking Arguments in Ethics

Most of the literature treats debunking arguments in a binary way: either the target moral beliefs are rendered unjustified or they aren't; either we endorse the premises of a debunking argument, or we don't. However there's a different, and more nuanced, way to see such arguments – as leading to the conclusion that we should become *less confident* in the target moral beliefs, and as relying on premises that we have varying degrees of confidence in.[1] This picture is suggested by Bayesian epistemology – our leading theory of normative constraints on probabilistic belief. There have only been a few mentions of Bayesianism in relation to debunking (Brosnan, 2011; Goldman, 2016).[2] I believe, however, that a thorough Bayesian analysis can further the debunking debate, and guide us in revising our beliefs in light of debunking. By clarifying key mechanics and uncovering crucial assumptions, such analysis will help us better understand the proper strength and scope of debunking arguments.

This is what I hope to achieve in this chapter, and some payoffs of my analysis are as follows: first, I highlight an important but implicit condition for debunking arguments to work – such arguments don't undermine everyone's confidence in the relevant moral beliefs; rather they undermine confidence only for agents who previously treated intuitions as evidence. Second, I argue that the most convincing kinds of debunking arguments will only conclude that we should reduce confidence to some degree, rather than that we should reduce confidence beyond some threshold, like 0.5. Third, I provide a method for determining how significant this reduction should be, in light of different kinds of evidence – about the kinds of epistemic

[1] I will assume that our confidence/credence in a moral belief is equal to our best estimate of its probability of truth. This links this chapter's Bayesian analysis with the Probability of Truth Project that I argued for in Chapter 2.

[2] Also see O'Neill (2015), who, in her analysis of debunking, cites Roush's (2007) conditional probability approach to truth-tracking. Goldman is actually concerned with debunking in metaphysics – as mentioned earlier, I will restrict myself only to talking about debunking arguments in ethics, but I believe the same mechanics will apply to debunking in any area of philosophy.

flaws that might be at play, about the different possible origins of our moral beliefs, and about the kinds of background normative assumptions that we're entitled to make. I also argue that one kind of debunking argument should have greater marginal impact on our credences than another – but that the empirical evidence necessary for supporting this stronger kind of argument is harder to come by. With this analysis, I hope to move the debunking debate from a coarse-grained description of epistemic impact – about whether our moral beliefs are justified, for instance – to a finer-grained and more nuanced treatment – about when we should reduce confidence in our moral beliefs, and by how much.

My discussion will be structured as follows: I introduce the Bayesian framework in section 6.1, and give some reasons for why it's well-suited for understanding debunking arguments. In section 6.2, I specify the possible hypotheses and evidence at play in the debunking debate. I then present the Bayesian model in section 6.3, draw out some implications and extensions of this model in section 6.4, and conclude in section 6.5.

6.1 The Bayesian Framework and Why It's Appropriate for Analysing Debunking

Bayesian epistemology is a normative theory of probabilistic belief – it concerns how we *should* revise our beliefs in light of the evidence. It takes agents to have doxastic attitudes called *credences* – these are modelled using numbers from 0 to 1, and indicate how confident an agent is in a proposition. For example, I might have 0.6 credence in the proposition "The patient has the disease". Bayesians propose normative constraints to govern our credences. One core set of constraints governs how credences should behave at a time – these are Kolmogorov's probability axioms.[3] Another constraint governs how credences should evolve over time as the agent gains more evidence. Let h be the hypothesis that the agent is entertaining (for instance, that "The patient has the disease"), let e be the evidence that she

[3] For a precise statement of these, see Talbott (2016).

gets (for instance, that "The patient returned a positive test result"), and assume for now that the evidence is learned for sure. Let $Cr_{new}(.)$ be the new, post-learning credence function, and $Cr_{old}(.)$ be the old credence function. Bayesians claim that the agent should update her credences according to Conditionalization, which says that

$$Cr_{new}(h) = Cr_{old}(h|e), given\ that\ Cr_{old}(e) > 0\ and\ Cr_{new}(e) = 1$$

Combining Conditionalization with Bayes Theorem and the Law of Total Probability, we get

$$Cr_{new}(h)$$
$$= Cr_{old}(h|e)$$
$$= \frac{Cr_{old}(e|h).\,Cr_{old}(h)}{Cr_{old}(e)}\ (from\ Bayes\ Theorem)$$
$$= \frac{Cr_{old}(e|h).\,Cr_{old}(h)}{Cr_{old}(e|h).\,Cr_{old}(h) + Cr_{old}(e|{\sim}h).\,Cr_{old}({\sim}h)}\ (from\ the\ Law\ of\ Total\ Probability)$$

In this equation, we have a way of relating the agent's new credence in the hypothesis, $Cr_{new}(h)$, to some of her old credences – $Cr_{old}(h)$, which is known as a *prior* for h, and $Cr_{old}(e|h)$ and $Cr_{old}(e|{\sim}h)$, which are known as *likelihoods*. The orthodox Bayesian picture has it that an agent starts off with some initial probability function – which includes these priors and likelihoods – and then updates via Conditionalization as she receives more evidence. In each update, she changes her credence in the evidence to 1, and then redistributes her credences proportionately to the hypothesis h and its alternatives.

Consider, for example, a doctor who is trying to determine whether a patient has a disease. (This is a stock example of Bayesian epistemology, but I believe it's quite analogous to the case of debunking – and so is worth going through in detail.) We start off in time period 1 (denoted by a credence function with subscript 1), where the doctor, informed by frequency data from the population, has some prior credences in hypotheses about the patient's condition. Let $Cr_1(disease)$ be the doctor's credence in the hypothesis that "The patient has

the disease", and $Cr_1(\sim disease)$ be the doctor's credence in the alternative hypothesis. Let's say that disease is quite rare in the population, so the doctor thinks it's more likely that the patient doesn't have the disease. We might assign the values for their credences as

$$Cr_1(disease) = 0.2$$

$$Cr_1(\sim disease) = 1 - Cr_1(disease) = 0.8$$

The doctor wants to administer a test for the disease. To update on the results of this test, they need the likelihoods – that is, the conditional credences that they get a positive result, given that the patient actually has, or actually doesn't have, the disease. Let $Cr_1(positive|disease)$ be the doctor's likelihood that the patient would return a positive test result, given that the patient actually has the disease. We can think of this as representing the doctor's beliefs about the true positive rate – that is, about the proportion of positive results that would be returned for patients who actually have the disease. Let $Cr_1(positive|\sim disease)$ be the doctor's likelihood that the patient would return a positive test result, given that the patient doesn't actually have the disease. We can think of this as representing the doctor's belief about the *false* positive rate – the rate at which the test gives a positive result, given that the patient *doesn't* actually have the disease. Given that the doctor sees the test as reliable, we might assign values for these likelihoods as follows:

$$Cr_1(positive|disease) = 0.9$$

$$Cr_1(positive|\sim disease) = 0.3$$

In time period 2, the doctor administers the test and obtains a positive test result. How should they update their credence in the hypothesis that the patient has the disease? To find out, we apply Conditionalization, Bayes Theorem, and the Law of Total Probability. Let $Cr_2(disease)$ be the doctor's period 2 credence that the patient has the disease – we can compute this using the formula

$$
\begin{aligned}
&Cr_2(disease) \\
&= Cr_1(disease|positive) \\
&= \frac{Cr_1(positive|disease).\,Cr_1(disease)}{Cr_1(positive)} \\
&= \frac{Cr_1(positive|disease).\,Cr_1(disease)}{Cr_1(positive|disease).\,Cr_1(disease) + \ Cr_1(positive|{\sim}disease).\,Cr_1({\sim}disease)}
\end{aligned}
$$

Substituting our values for $Cr_1(disease)$, $Cr_1(positive|disease)$, and $Cr_1(positive|{\sim}disease)$, we have

$$
\begin{aligned}
&Cr_2(disease) \\
&= \frac{0.9 \times 0.2}{0.9 \times 0.2 + \ 0.3 \times 0.8} \\
&= 0.43 \ (to \ 2 \ d.p.)
\end{aligned}
$$

After receiving the positive test result in stage 2, the doctor should adopt a credence of 0.43 in the hypothesis that the patient has the disease. This is an increase from $Cr_1(disease)$, which was 0.2.

This example is a good illustration of how Bayesians might prescribe changes in credences, given certain values for the priors and likelihoods. But Bayesians might impose further constraints on these values too. For instance, Sober (2008, pp. 24–30) argues that in many scientific endeavours, scientists can use past frequency data to inform their likelihoods. In our example, the likelihood $Cr_1(positive|disease)$ could be informed by giving the test to people whom we already know have the disease, and seeing how many positive results are obtained.

These Bayesian constraints are often taken to govern an agent's credences as they should change over time, but there's another interpretation that fits better with my project. Marc Lange (1999) argues that Bayesian norms can also govern the steps in the arguments by which our current credences are justified. He gives the example of a weather forecaster who wants to argue that her current 0.8 credence in 'It will rain and not snow tomorrow' is justified. She

starts off with a prior probability distribution over this hypothesis and its alternatives – a distribution that other scientists will grant her as an argumentative 'free move' that needs no further justification. She then introduces the first piece of evidence – say, that 'Today's barometric pressure is 29 torr' – and 'updates' her initial distribution to an intermediate conclusion (another probability distribution) via Conditionalization. She then continues introducing further pieces of information that she has, 'updating' in this way until she has considered all the relevant evidence. If her justificatory argument works, then she should eventually arrive at a distribution that assigns 0.8 credence to 'It will rain and not snow tomorrow'. If she doesn't, then something has gone wrong with her argument, and she is not justified in assigning 0.8 credence in the hypothesis (Lange, 1999, pp. 302–306).

On this interpretation, it's not as though her actual credences change as she considers each new piece of evidence – rather, each intermediate probability distribution represents an intermediate conclusion on the way to justifying her actual credence. Furthermore, the priors represent an argumentative 'free move' that can be appealed to without further justification, while the likelihoods represent the correct evidentiary weight she believes should be accorded to each piece of evidence. Lange uses this story to account for theory confirmation in the sciences, but I believe it can be usefully applied to philosophical arguments too. The debunkers' opponents cite some intuitions or beliefs as evidence to support their credence in moral hypotheses – much like how the forecaster cites the barometer reading as evidence that it will rain and not snow tomorrow. In response, the debunkers argue that given some debunking story, these intuitions or beliefs shouldn't be accorded much evidential weight, if any at all – so their opponents are irrational in assigning the credence they do. That is, these debunkers view their opponents in the way a forecaster might view their colleagues when realizing that a barometer is faulty. This forecaster would argue that we shouldn't give much evidentiary weight to the barometer's readings, and that anyone who did so would end up with an irrational credence. The Bayesian picture is thus seems well-equipped to model the debate between the debunkers and their opponents, and I hope to demonstrate the benefits of reconceiving the debate in Bayesian terms below.

6.2 Debunking in Bayesian Terms

Bayesianism works with some hypothesis (like "The patient has the disease") and evidence (like "The patient returned a positive test result"). What might these be in the debunking debate? First, let's look at some prominent hypotheses:

- Categorical moral reasons exist – these are reasons that apply to everyone regardless of what their desires are, and have particularly inescapable force (Joyce, 2006; Morton, 2016)
- It is morally permissible to redirect the trolley in the Switch, Footbridge, and Loop cases (Liao et al., 2012; Machery, 2017, Chapter 2)
- Cooperation is morally good (Brosnan, 2011, p. 53)
- "The fact that something would promote the interests of a family member is a reason to do it", "We have greater obligations to help our own children than we do to help complete strangers" (Street, 2006, p. 115)

I want the later analysis to apply to these different possibilities, so I'll just work with a generic moral hypothesis h. What about the evidence? Here, too, there is variation. We might take the evidence for a moral proposition that h to be that

- Everyone believes that h (Parfit, 1986, p. 186),
- You believe that h (Brosnan, 2011; Goldman, 2016),
- You have the intuition that h (Huemer, 2005)

Here I cannot remain similarly neutral. First, treating the evidence as everyone's believing that h seems to be shortchanging the debunking argument. As discussed earlier, the debunkers don't just want to remove the support rendered by widespread agreement of others – they also want to remove the evidential support created by an individual's own moral beliefs or intuitions. After all, they take themselves to have debunked our moral beliefs, even in cases where people disagree about moral matters. So while they might be trying to undermine widespread agreement, that cannot be all that they are doing.

Secondly, we might take the evidence to be your believing that h. There are some issues with this. Normatively speaking, we generally shouldn't treat our beliefs as evidence, since that seems like double-counting. Moreover, as a descriptive matter, we don't usually treat our beliefs as evidence either. As White (2010, p. 585) argues,

> I don't think over whether p and then upon coming to the conclusion that p think 'p, and now I believe that p. Smart fellow that I am it is unlikely that I would believe that p if it wasn't true. So this is further evidence that p.' and thereby increase my confidence that p.

But perhaps things are different in the moral case: we might have been handed down certain beliefs from our culture and evolutionary history, and just accepted that there were good arguments to be made for these beliefs, even if we didn't know what those arguments were. This might supply reasons for treating our beliefs as evidence, and it also might be a good descriptive picture of how we arrived at our actual moral beliefs. While I have some sympathy for this way of understanding the evidence, I believe that modelling the situation using our intuitions as evidence (see below) is the least controversial route. I believe, however, that the same Bayesian machinery I propose can be fitted to using beliefs as evidence too (see n.7).[4]

Third, and most plausibly, we can take our intuitions to be the evidence.[5] We often seem to make the following inference: "I have the intuition that h. This is evidence that h, and I should increase my credence in h." For example, if I have the intuition that it is morally permissible to redirect the trolley toward the one person in the Loop case, I will often increase my credence that it is indeed permissible to do so. Or when I have an intuition that I have a categorical moral reason to perform this or that action, regardless of what my desires are, I would also increase my credence that such reasons exist. We think we should sometimes update our credence in the moral hypothesis in this way – and, as a descriptive matter, we do sometimes update like this. So I believe the Bayesian model should take your having certain intuitions as the evidence. We can be neutral about what exactly intuitions are – they could be dispositions to believe, intellectual seemings, or some other sui generis mental state. What's crucial here is that the intuitions are separate from the final credence in h – in the sense that

[4] Thanks to Katie Steele and Kim Sterelny for the helpful discussion here.

[5] Also see Climenhaga (2018), who argues that philosophers do indeed use intuitions as evidence in philosophy.

our final credence in h could be low, even when we have the intuition that h (and vice versa when we lack the intuition that h).

6.3 The Bayesian Model of Debunking

Having clarified the evidence and hypotheses, I'll now present the Bayesian model, which involves three stages. Stage 1 defines the prior credence in the moral hypothesis h and its alternatives, and some of the likelihoods at play. In stage 2, the debunker's opponents introduce some intuitions as evidence. Updating on these intuitions increases our credence in the moral hypothesis. In stage 3, the debunker introduces some empirical evidence that reveals the dubious origins of these intuitions. This changes the evidentiary weight accorded to these intuitions, mandating a reduction from the stage 2 credence in the moral hypothesis.

6.3.1 Stage 1

As established earlier, we'll work with a generic moral hypothesis h, and with the evidence being your having the intuition that h. Remember that the priors represent an argumentative 'free move' – a probability distribution over the hypothesis and its alternatives, which both the debunkers and their opponents agree needs no further justification. What might this distribution look like? One plausible answer might be that h is as likely to be true as it is to be false,[6] so we can assign the priors as follows:

$$Cr_1(h) = 0.5$$

$$Cr_1(\sim h) = 1 - Cr(h) = 0.5$$

Next, the likelihoods: remember these represent the evidentiary weight we believe should be accorded to the evidence – which, in our case, is your having the intuition that h. Let

[6] This follows the principle of indifference, which says that "Given n mutually exclusive and jointly exhaustive possibilities, none of which is favored over the others by the available evidence, the probability of each is 1/n." (Weisberg, 2017, sec. 2.1) There are various problems with this principle, but we can set them aside for the purposes of this chapter. Some philosophers want to use intuitions to directly inform the priors, but I prefer to draw out the epistemic influence of intuitions using an explicit updating procedure, as I detail in stages 2 and 3 below.

Cr_1(int|h) be the likelihood that you have the intuition that h, given that h is true, and Cr_1(int|~h) be the likelihood that you have the intuition that h, given h is false.[7] If we grant for now that intuitions are good indicators of the truth, we accord them significant evidentiary weight. This can be represented by a high Cr_1(int|h) and a low Cr_1(int|~h), as follows:

$$Cr_1(int \mid h) = 0.9$$

$$Cr_1(int \mid \sim h) = 0.1$$

These likelihoods are analogous to the ones in section 6.1, which represent the doctor's beliefs about the true positive rate and the false positive rate of the test results. In the same way, we can see our intuitions as test results or instrument reports that indicate something about the moral hypothesis – these intuitions likewise yield true and false positives at some rate.[8]

I have used exact values for the priors and likelihoods to illustrate the model here, but we shouldn't take these values too seriously. We could instead also perform a robustness analysis, over a range of different possible values for the priors and likelihoods, in more specific cases.

6.3.2 Stage 2

In stage 2, we introduce the evidence that you have the intuition that h. Applying Bayes Theorem and the Law of Total Probability, the new credence we should adopt in the moral hypothesis h is

$$Cr_2(h)$$
$$= Cr_1(h|int)$$
$$= \frac{Cr_1(int|h).\,Cr_1(h)}{Cr_1(int|h).\,Cr_1(h) + \; Cr_1(int|\sim h).\,Cr_1(\sim h)}$$

[7] If we instead took our beliefs as evidence, these likelihoods will represent the probability that we would have ended up with the beliefs we do, given the truth or falsity of the moral hypothesis. On this understanding, we are not updating on our intuitions, but rather working backward to see if there was some good argument supporting our beliefs or not, given what we know about the origins of such beliefs.

[8] See O'Neill (2015), who also talks about some debunking arguments in the language of true and false positives.

Substituting our values for $Cr_1(h)$, $Cr_1(\sim h)$, $Cr_1(int|h)$ and $Cr_1(int|\sim h)$, we have

$$Cr_2(h)$$
$$= \frac{0.9 \times 0.5}{(0.9 \times 0.5) + (0.1 \times 0.5)}$$
$$= 0.9$$

Call this new credence, $Cr_2(h)$, the *evidential credence*.[9] It represents the confidence we should have in the moral hypothesis h, after updating on having the relevant intuition. Just like how the doctor takes a positive test result as indication of the disease, the moral philosopher takes their having the intuition as evidence for the moral hypothesis. We arrive at an evidential credence of 0.9, an increase from the prior credence of 0.5.

6.3.3 Stage 3

The debunker now introduces a causal story about the intuitions that were treated as evidence – call this *story*. For instance, the debunker might allege that your intuitions were produced by evolutionary processes that aim to promote reproductive success, rather than to track the moral truths. This is supposed to be analogous to the doctor discovering that their test results are unreliable – I'll flesh out more details of the debunking story later on. For now, notice that in stage 3, we have two pieces of evidence – *int* and *story*. So the stage 3 credence in h should be

$$Cr_3(h)$$
$$= Cr_1(h|int\ \&\ story)$$

[9] Here I'm following the terminology of Kotzen (forthcoming), who talks about the credences involved in defeat.

Call this new credence the *debunked credence* – it tells us how confident we should be in h, after learning both the intuition and the debunking story. To relate this debunked credence to easily interpretable terms, I'll expand it in a slightly different way from before, such that

$$
\begin{aligned}
&Cr_3(h)\\
&= Cr_1(h|int\ \&\ story)\\
&= \frac{Cr_1(h\ \&\ int\ \&\ story)}{Cr_1(int\ \&\ story)}\ (from\ Ratio\ Formula)\\
&= \frac{Cr_1(h\ \&\ int\ \&\ story)}{Cr_1(int\ \&\ story\ \&\ h) + Cr_1(int\ \&\ story\ \&\ {\sim}h)}\ (from\ expansion\ of\ denominator)\\
&= \frac{Cr_1\big(int\ \big|\ story\ \&\ h\big).Cr_1(h|story).Cr_1(story)}{Cr_1\big(int\ \big|\ story\ \&\ h\big).Cr_1(h|story).Cr_1(story) + Cr_1\big(int\ \big|\ story\ \&{\sim}h\big).Cr_1({\sim}h|story).Cr_1(story)}\\
&= \frac{Cr_1\big(int\ \big|\ story\ \&\ h\big).Cr_1(h|story)}{Cr_1\big(int\ \big|\ story\ \&\ h\big).Cr_1(h|story) + Cr_1\big(int\ \big|\ story\ \&{\sim}h\big).Cr_1({\sim}h|story)}
\end{aligned}
$$

Line 3 follows from the Ratio Formula, except now for a conditional credence that conditions on both pieces of evidence. Line 4 expands the denominator Cr_1(int&story) into the sum of credences in the two possibilities, Cr_1(int&story&h) and Cr_1(int&story&~h). Line 5 expands all terms in the numerator and denominator using the chain rule for conditional credences.[10] Line 6 factors out Cr_1(story) from the numerator and denominator – arriving at a formula for the debunked credence that uses the priors and some new conditional credences, Cr_1(int|story&h), Cr_1(int|story&~h), Cr_1(h|story), and Cr_1(~h|story).[11]

The debunker's challenge imposes constraints on these new conditional credences – this then affects the value of the debunked credence. Let's start with the latter two, Cr_1(h|story) and Cr_1(~h|story). These represent the credence we should adopt in h and ~h respectively, given that we learn only about *story*, but not about having any relevant intuitions. It seems that learning *story* should affect our credence in h only *through* undermining the intuitions that

[10] The chain rule says that Cr(A&B&C) = Cr(A|B&C).Cr(B|C).Cr(C)

[11] This same kind of derivation is used by Bovens and Hartmann (2004, p. 70) in modelling the reliability of scientific instruments.

were used as evidence. So if we just learned *story* alone without learning about having the relevant intuitions, this shouldn't have any effect on our credence in the moral hypothesis at all. This is analogous to learning about the reliability of a medical test instrument, before we have heard about or updated on any results from that instrument – in such a case, it seems like our opinions about the patient's condition shouldn't change.[12] Similarly, in the debunking case, the following two conditions should apply

$$Cr_1(h|story) = Cr_1(h)$$

$$Cr_1(\sim h|story) = Cr_1(\sim h)$$

That is, when we learn *story* alone, our credence in h should remain equal to our prior credence in h; likewise with our credence in not-h. These mirror similar conditions for undercutting defeaters (Kotzen, forthcoming, p. 10) – which bodes well for the model, since debunking arguments are often likened to undercutting defeaters (Kahane, 2011, p. 106; McGrath, 2014, pp. 210–211; Lutz, 2018). Substituting these conditions into our formula for the stage 3 debunked credence, we have

$$Cr_3(h) = \frac{Cr_1\big(int \,\big|\, story \,\&\, h\big).\, Cr_1(h)}{Cr_1\big(int \,\big|\, story \,\&\, h\big).\, Cr_1(h) + \; Cr_1\big(int \,\big|\, story \,\&\sim h\big).\, Cr_1(\sim h)}$$

And we can compare this with the evidential credence from stage 2,

$$Cr_2(h) = \frac{Cr_1(int|h).\, Cr_1(h)}{Cr_1(int|h).\, Cr_1(h) + \; Cr_1(int|\sim h).\, Cr_1(\sim h)}$$

Notice that what's changed is the likelihoods in these two formulas. The stage 2 evidential credence was computed with Cr_1(int|h) and Cr_1(int|~h); in the stage 3 debunked credence, these are replaced by Cr_1(int|story&h) and Cr_1(int|story&~h) respectively. The debunker can be interpreted as constraining Cr_1(int|story&h) and Cr_1(int|story&~h), in order to achieve a

[12] For analogous cases and conditions, see Kotzen (forthcoming, p. 10) and Bovens and Hartmann (2004, p. 58).

reduction in credence as we move from stage 2 to 3. To understand how this works, we'll have to delve deeper into the details of the debunking story.

6.3.4 Two Debunking Stories, Two Constraints on Likelihoods

Recall that there were two epistemic flaws associated with counterfactual reliability – chanciness and inevitability. I believe these flaws best capture the two different debunking stories that we can tell about our moral intuitions too. Consider chanciness first. The debunking story here can be understood as alleging that we could easily have had different moral intuitions, even when the moral facts remained the same. To revisit an earlier example, Joyce (2006, 2016) could also be read as arguing that it was evolutionarily lucky for us to have the moral intuitions that we do. Given the contingencies of the evolutionary process, it could easily have been that we had different moral intuitions, even though the moral facts remained the same. Challenges concerning demographic variation in our moral intuitions, or variations across different frames, phrasings, or conditions of a thought experiment work in roughly the same way: the moral facts remain the same across different frames and conditions, and across demographic variables like culture, yet our moral intuitions vary (Liao et al., 2012; Machery, 2017, Chapters 2–3; Sinnott-Armstrong, 2011). We can read all these challenges as highlighting the low rate of true positives, given some debunking story – in particular, as alleging the low probability of your having the intuition that h, given that h and story are true. Call this kind of epistemic impact *true positive less likely*. This can be modelled by lowering $Cr_1(\text{int}|\text{story}\&h)$ to 0.5 – as compared to $Cr_1(\text{int}|h)$, which was 0.9. Let's assume that the other likelihood, $Cr_1(\text{int}|\text{story}\&\sim h)$, remains the same as $Cr_1(\text{int}|\sim h)$ at 0.1. Then constraining $Cr_1(\text{int}|\text{story}\&h)$ to 0.5 means we should adopt a debunked credence of 0.83 in h – a reduction from stage 2's evidential credence of 0.9 in h. By showing that the true positive rate for our intuitions is lower than what we initially thought, the debunker achieves a reduction

in credence in h from stage 2.[13] More generally, for *true positive less likely* to work, the debunkers need to ensure that

$$Cr_1(int \mid story \,\&\, h) < Cr_1(int|h)$$

The second flaw, inevitability, is employed in a different kind of debunking story. The idea here is rather that we would have still have the same intuitions, even if the moral facts were changed. For example, Morton (2016, p. 242) argues that evolutionary processes would have made us believe that categorical moral reasons exist, whether or not such reasons actually did exist. This challenge can be read as alleging the high rate of *false* positives, given some debunking story – in particular, it highlights the high probability that you have intuition that h, given that h is *false* and story is true. Call this kind of epistemic impact *false positive more likely*, since false positives are shown to be more likely than you thought. In terms of credences, this might be modelled as a constraint increasing Cr_1(int|story&~h) to 0.5 – as compared to Cr_1(int|~h), which was 0.1. To measure this impact, let's suppose that the other likelihood, Cr_1(int|story&h), remains equal to Cr_1(int|h) at 0.9. Then constraining Cr_1(int|story&~h) to 0.5 would mean we should adopt a debunked credence of 0.64 in h – a significant reduction from the 0.9 evidential credence in stage 2. By showing that the false positive rate for intuitions is higher than we initially thought, the debunkers also achieve a reduction of credence in h from stage 2.[14] Generally, for *false positive more likely* to work, the debunkers need to show that

$$Cr_1(int \mid story \,\&\, \sim h) > Cr_1(int|\sim h)$$

[13] The impact of *true positive less likely* can be used to model debunking arguments from contingency, accidentality or lack of safety (Bogardus, 2016, pp. 645–647; Handfield, 2016, p. 68; Joyce, 2006, p. 181), and from counterfactual disagreement (Bogardus, 2016, pp. 655–657).

[14] The impact of *false positive more likely* can be used to model debunking arguments from insensitivity (Bogardus, 2016, pp. 638–640; Morton, 2016, pp. 235, 242–243; Ruse & Wilson, 1986, pp. 186–187), and from the screening-off of evidence (White, 2010, pp. 580–581). Climenhaga (2018, n. 26) also uses this kind of impact to model the practice of offering error theories of our intuitions in philosophy.

Either of the above two constraints – concerning *true positive less likely* or *false positive more likely* – will suffice for reducing confidence in h as we move from stage 2 to 3. The formula for the debunked credence $Cr_3(h)$ also enables us to quantitatively estimate the epistemic impact of both these constraints working *together*. This represents an advance, since these impacts are usually considered in isolation.[15] Moreover, notice that this Bayesian model doesn't commit us to saying that the debunked credence must go below some specific threshold (like 0.5), as many debunking arguments seem to conclude.[16] Rather, it says we should reduce confidence by some amount, and leaves open what the magnitude of this reduction should be.[17]

6.4 Implications and Extensions of the Bayesian Model

Having seen how the general model works, let's look at some implications and extensions.

6.4.1 Debunking effect only conditional on updating erroneously

The debunkers can be thought of as arguing that our credence in the moral hypothesis should be reduced as we move from stage 2 to stage 3. As seen earlier, to mandate this reduction in confidence, they need at least one of the following conditions to hold:

$$Cr_1(int \mid story \,\&\, h) < Cr_1(int|h)$$

$$Cr_1(int \mid story \,\&\, \sim h) > Cr_1(int|\sim h)$$

[15] Debunking arguments alleging the insensitivity of our moral beliefs are often run separately from those alleging their lack of safety, for instance.

[16] For instance, Joyce concludes that evolutionary considerations render our moral beliefs "unjustified" (2006, p. 180), that we should "cultivate agnosticism" (2006, p. 181) about our moral beliefs until we get further evidence, that we should keep "an open mind about whether there exists anything that is morally right or wrong" (2006, p. 181). Morton concludes that our moral beliefs are "defeated" (2016, p. 242) by learning about their evolutionary origins, and that we "would be no longer justified" (2016, p. 243) in holding them without further independent justification. Street (2006, p. 125) goes for the stronger conclusion that our moral beliefs are "likely to be false". Machery (2017, p. 95) argues that unless we know that a specific case is an exception, we "ought to suspend judgment" about situations described in various philosophical cases – including moral ones – because of the demographic and presentation effects found.

[17] Joyce (2016, p. 125) moves closer to this approach when he says that our confidence in our moral beliefs should be "dented".

Notice, however, that these conditions depend crucially on the values of $Cr_1(int|h)$ and $Cr_1(int|{\sim}h)$ that were used for updating in stage 2. For instance, upon learning about the debunking story, you should reduce credence in the moral hypothesis only if you previously updated with a $Cr_1(int|h)$ that was higher than the constrained $Cr_1(int|story\&h)$. That is, you should reduce credence in the hypothesis only if you updated on the intuitions thinking that the true positive rate was higher than it actually is. And vice versa with $Cr_1(int|{\sim}h)$ and the false positive rate – you should only reduce confidence in h if you updated thinking that the false positive rate was lower than it actually is. This highlights the essentially historical nature of debunking arguments – such arguments don't dictate that *every* agent should reduce confidence in h. Rather, these arguments are directed only to agents who updated on their intuitions using an erroneous evidentiary weight. If someone updated on them with the correct evidentiary weight, then the debunker's constraints shouldn't affect their credence in the moral hypothesis at all.[18] Thus the Bayesian model makes clear that the debunking argument assumes a descriptive claim too: it assumes that we have previously updated on our intuitions with an erroneous evidentiary weight.

Recognising this essentially historical nature also reveals how constraints on $Cr_1(int|story\&h)$ and $Cr_1(int|story\&{\sim}h)$ could have a perverse effect for the debunker. Suppose someone initially updated their beliefs with a very low evidentiary weight on their intuitions – taking $Cr_1(int|h)$ to be 0.001, for instance. The debunker then introduces *story* as evidence, which constrains $Cr_1(int|story\&h)$ to be 0.5. This person should now give their intuitions *more* evidentiary weight than before, since true positives are more likely than they initially thought. Other things being equal, the debunker's constraint will dictate that this person should *increase* credence in the moral hypothesis.[19] This starkly illustrates how the debunking

[18] See also Climenhaga (2018, p. 86), who argues that if a philosopher didn't take intuitions as evidence for their theory, they shouldn't have to care at all about debunking explanations of those intuitions. But note the perverse effect that I detail in the next paragraph.

[19] This might be what we should do in light of discovering surprising agreement or robustness in our beliefs/intuitions, as argued by Knobe (2019).

argument's impact depends on what evidentiary weight we initially assigned to our intuitions in stage 2.

6.4.2 Special case #1: Where Cr1(int|story&h) = Cr1(int|story&~h)

In the basic model, we looked at constraints on Cr_1(int|story&h) and Cr_1(int|story&~h), and found that a constraint on either is sufficient to reduce our credence in h. There are some special cases of epistemic impact that are worth highlighting too. Consider, for instance, the case where the likelihoods are constrained such that

$$Cr_1(int \mid story\ \&\ h) = Cr_1(int|story\&\sim h)$$

In this special case, it's equally likely that you have the intuition that h, given that h and *story* are both true, as it is given that h is *false* and *story* is true. Here, the intuition is no guide to the moral hypothesis at all, and we should adopt a stage 3 debunked credence that's *equal* to our prior credence in h from stage 1. To illustrate, let Cr_1(int|story&h) = Cr_1(int|story&~h) = x, then the debunked credence is

$$Cr_3(h)$$
$$= \frac{Cr_1(int \mid story\ \&\ h).\ Cr_1(h)}{Cr_1(int \mid story\ \&\ h).\ Cr_1(h) + \ Cr_1(int \mid story\ \&\sim h).\ Cr_1(\sim h)}$$
$$= \frac{0.5x}{0.5x + \ 0.5x}$$
$$= 0.5 \ = Cr_1(h)$$

Brosnan (2011) is chiefly concerned with this special case, where the debunking brings us back to our prior credence. But we should also think about weaker epistemic impacts – for as long as either of the two conditions for *true positive less likely* and *false positive more likely* are satisfied, the debunking argument will have some effect on our moral credences.

6.4.3 Special case #2: Where Cr(int|story&h) < Cr(int|story&~h)

Another special case involves constraining the likelihoods such that

$$Cr_1(int \mid story\ \&\ h) < Cr_1(int|story\&{\sim}h)$$

That is, the constraints make it more likely that you have the intuition that h, given h is *false* and *story* is true, than it is that you have the intuition, given that h and *story* are both true. Here, learning about the debunking story shows us that the intuition is an anti-reliable indicator of the moral hypothesis – the intuition becomes evidence for ~h! In such cases, the stage 3 debunked credence in h could go *below* the stage 1 prior credence in h.[20] For instance, if Cr_1(int|story&h) = 0.3, Cr_1(int|story&~h) = 0.8, then the debunked credence in h will be 0.27 (a great reduction from the prior credence of 0.5).

$$Cr_3(h)$$
$$= \frac{Cr_1(int \mid story\ \&\ h).\ Cr_1(h)}{Cr_1(int \mid story\ \&\ h).\ Cr_1(h) + \ Cr_1(int \mid story\ \&{\sim}h).\ Cr_1({\sim}h)}$$
$$= \frac{0.3\ \times 0.5}{(0.3\ \times 0.5) + \ (0.8\ \times 0.5)}$$
$$= 0.27 \ll \ Cr_1(h)$$

Brosnan (2011, p. 54) actually counts this case as one of tracking success – since our moral intuitions are indicators, albeit anti-reliable ones, of the moral facts. I disagree – it would be cold comfort to learn that we should conclude exactly the opposite of what our moral intuitions tell us.

Some have also argued that the debunking story cannot achieve this effect. They claim that such stories can at best reduce the support rendered by our intuitions, but cannot show such intuitions to be anti-reliable (White, 2010, p. 590). I'll remain neutral about whether our

[20] If the debunking effect works this way, then *story* becomes what Kotzen (forthcoming, p. 10) calls a bidirectional defeater.

intuitions can be shown to be anti-reliable or not – I only offer the Bayesian story as a way of modelling such an impact.

6.4.4 Marginal Effects of Constraining False Positive vs True Positive Rate

The model also has implications for the relative size of the two kinds of epistemic impact. It seems that point for point, *false positive more likely* has greater marginal impact on the debunked credence in h than *true positive less likely*. As an illustration, refer to the initial examples from section 3. The example of *false positive more likely* constrained $Cr_1(int|story\&\sim h)$ to a value that's 0.4 greater than $Cr_1(int|\sim h)$ – a 0.4 increase in the false positive rate. This lowered the debunked credence to 0.64. On the other hand, the example of *true positive less likely* constrained $Cr_1(int|story\&h)$ to a value that's 0.4 lower than $Cr_1(int|h)$ – a 0.4 decrease in the true positive rate. This only lowered the debunked credence to 0.825.

This difference in marginal impact – where an instance of *false positive more likely* has greater impact on our credence in the moral hypothesis than an equal-sized instance of *true positive less likely* – is robust across all values that are relevant for the debunking debate. In particular, *false positive more likely* has greater marginal impact as long as the intuition is taken as some positive indication of the hypothesis, and our prior credence in the hypothesis lies between 0 and 1 (see Appendix for proof).

6.4.5 The Empirical Evidence for True and False Positives

So far, we've just assumed that the debunker manages to constrain the likelihoods $Cr_1(int|story\&h)$ and $Cr_1(int|story\&\sim h)$. But we might ask *how* they could achieve this – in particular, how their evidence could constrain the relevant likelihoods. The debunker might draw inspiration from our earlier example of the doctor using the test results to determine whether their patient has the disease. Recall that the doctor needed to assign the values for the likelihoods $Cr_1(positive|disease)$ and $Cr_1(positive|\sim disease)$, which are the likelihoods that we

obtain a positive result, given that the patient actually has or doesn't have the disease respectively. We entertained a frequency-based strategy for doing so: supposing we wanted to set the credence Cr_1(positive|disease), we could give the test to patients whom we already know have the disease, and see what proportion of positive results are returned. Could the debunker use this same strategy to constrain their likelihoods?

There are some issues here, since it's unclear whether the moral hypothesis h is true in the actual world, where our evidence comes from. Given this uncertainty, it's unclear which likelihood will be constrained by our evidence. The moral case is thus more like the case of a doctor who has many test results from many patients, but doesn't know for sure if any specific patient actually has the disease or not. But perhaps the debunker can still make some progress. For instance, they might concede for the sake of argument that h is true in the actual world – then the distribution of peoples' intuitions in the actual world can constrain Cr_1(int|story&h), leading to an impact of *true positive less likely*.

Things get a little more complicated for Cr_1(int|story&~h) and *false positive more likely* – which falls prey to roughly the same problems I raised in Chapter 4. It doesn't seem legitimate for the debunkers to just assume that h is false in the actual world. So any evidence for Cr_1(int|story&~h) will instead come from a possible world where h is false – and it seems that actual-world frequencies, and actual-world evidence, will just be unhelpful here. Moreover, if the moral truths are metaphysically necessary, the investigation constraining Cr_1(int|story&~h) will require looking at metaphysically impossible worlds, which creates a further disanalogy with the doctor case. The debunkers thus need to say more about how the empirical evidence can generate an instance of *false positive more likely*.

In the absence of good empirical information informing the likelihoods, it might be rational to use the principle of indifference to assign our credences in Cr_1(int|story&h) and Cr_1(int|story&~h).[21] This might help the debunker's argument. For instance, we might think it equally likely that we have an intuition, given that story and h are true, as it is that we don't

[21] Thanks here to Katie Steele.

have that intuition, given that story and h are true. Then $Cr_1(int|story\&h) = Cr_1(\sim int|story\&h) = 0.5$, which generates an instance of *true positive less likely*. If there are a range of different possible intuitions one could have about a case – and hence more partitions in the probability space – then $Cr_1(int|story\&h)$ could go even further below 0.5.[22] Of course, going this route means that the debunkers aren't really using empirical evidence to constrain the likelihoods – but rather only relying on the rational norm of the principle of indifference.

6.4.6 The Role of Priors

The prior credence in h, $Cr_1(h)$, plays an important role – even if the debunkers successfully constrain $Cr_1(int|story\&h)$ and $Cr_1(int|story\&\sim h)$, this prior credence still greatly influences the debunked credence. Brosnan (2011, pp. 54–55) points out that no value of the likelihoods will be sufficient for determining the debunked credence, and concludes that the debunkers aren't entitled to claiming that the moral hypothesis is likely false.

While that is correct, I wonder if it's a strong reply. Firstly, recall that we're interpreting the priors as an argumentative 'free move' that is allowed by both sides of the debate. The debunkers probably won't allow the 'free move' of assigning a high prior credence to the moral hypothesis, especially when this prior shouldn't be informed by any intuitions (whose evidential impact is taken into account later in the updating process). Even if they did allow assigning a high prior, all the work of defending against debunking is done merely by this assignment, rather than by any substantive argument. Leaning on the priors, as Brosnan does, merely stipulates the problem of debunking away, rather than giving a positive answer to it. Secondly, even if the debunkers don't show that our moral beliefs are likely false, they might still secure some weaker epistemic impact – for instance, by concluding that we should reduce confidence in these beliefs. Finally, just constraining the likelihoods still counts as a significant achievement for the debunkers. This undermines at least some of the evidence

[22] This kind of odds-based argument is suggested by Street (2006) at points – see also Shafer-Landau (2012, pp. 10–14).

cited by their opponents, and constrains the range of possible prior credences that could still vindicate our moral beliefs (that is, by still leading to a high debunked credence).

6.4.7 Constraining (h|story) and (~h|story)

In deriving the debunked credence, I said that the debunkers should argue for the following two conditions:

$$Cr_1(h|story) = Cr_1(h)$$

$$Cr_1(\sim h|story) = Cr_1(\sim h)$$

Recall that these concern how we shouldn't change our credence in the moral hypothesis h if we just learned about the debunking *story* alone, without learning about the intuition. Opponents of debunking will contend, however, that these conditions don't hold. Brosnan (2011, pp. 43, 59–61) argues that if we make some background normative assumptions, then some non-moral facts obtaining might raise the probability of some moral facts obtaining. So it could be that Cr_1(h|story) ≠ Cr_1(h); likewise that Cr_1(~h|story) ≠ Cr_1(~h). For example, if we assume that wellbeing is generally morally good, then it might be that a belief's being reproductively advantageous would raise the probability of its being true (as opposed to if it wasn't reproductively advantageous). This is because what promotes reproductive fitness will also generally promote wellbeing, which is itself a morally good thing. More generally, any third-factor account could support denying the above two conditions, since they postulate a third factor to explain how our evolutionarily-produced moral beliefs are correlated with the moral facts. To revisit a few other examples: Enoch (2010, pp. 430–432) argues that evolutionary causes tend to produce beliefs that promote survival, and survival tends to be a morally good thing. Wielenberg (2010, pp. 449–450) contends that for an organism to form moral beliefs about rights, it must have had sufficiently sophisticated cognitive faculties to do so. But if an organism had such faculties, it too would possess rights of its own. Copp (2008,

pp. 198–202) argues that morality has the function of helping societies to meet their basic needs of continued existence, and peace and cooperation. He then argues that evolutionary causes – both biological and cultural – would have produced moral beliefs that were approximately true, since the beliefs produced would have helped societies to meet these basic needs.

My Bayesian analysis can also contribute to this debate over what background assumptions are legitimate in the context of debunking. Recall our initial formula for the debunked credence, before substituting in the two constraints from above – this formula took $Cr_1(h|story)$ and $Cr_1(\sim h|story)$ as inputs. The debunker might concede to their opponents that $Cr_1(h|story) > Cr_1(h)$, for instance, but argue that $Cr_1(h|story)$ is only a tiny bit higher than $Cr_1(h)$. In fact, the opponents of debunking typically admit that the correlation between the moral and non-moral facts could be quite weak (Enoch, 2010, p. 430), or just that there *might* be a correlation (Brosnan, 2011, p. 61). In light of this, the debunker might just allow that the correlation exists, use that to compute $Cr_1(h|story)$ and $Cr_3(h)$, and argue that this correlation isn't strong enough to support high credence in the relevant moral propositions. Alternatively, the opponents of debunking can show how $Cr_1(h|story)$ and $Cr_1(\sim h|story)$ could in fact vindicate high moral credences. More generally, the formula for the debunked credence shows how we can move beyond a binary debate – about whether some non-moral facts raise the probability of a moral fact obtaining – to a graded one, about *how much* probability-raising is required to maintain a high credence in the relevant moral hypothesis. This formula also shows how we might use our credences in background normative assumptions – which might be from our metaethical theories, for instance – in computing our credences about moral hypotheses.

6.4.8 Uncertainty about story

In the basic model, we considered cases where we learned a single debunking story with certainty. However, the reality of the debunking debate is more complicated. In particular, our credences might be split between diverse debunking and non-debunking stories about

our moral intuitions – particularly when it comes to evolutionary debunking (Isserow, 2018, secs. 3–5). The Bayesian model can handle this too. We can use Jeffrey conditionalization, which tells us how we should update our credences in the moral hypothesis when our credences in the evidential statements change, but are not raised to certainty:

$$Cr_{new}(h) = Cr_{old}(h|e).Cr_{old}(e) + Cr_{old}(h|{\sim}e).Cr_{old}({\sim}e)$$

$$(given\ that\ 0 > Cr_{old}(e), Cr_{old}(h) > 1)$$

When applied to credences split between a debunking and non-debunking story, we get a formula like

$$Cr_{new}(h) = Cr_{old}(h|debunking\ story\ \&\ int).Cr_{old}(debunking\ story\ \&int) + Cr_{old}(h|nondebunking\ story\ \&\ int).Cr_{old}(nondebunking\ story\ \&int)$$

In effect, the formula tells us to weight the posterior credences – conditional on the debunking and non-debunking stories respectively – by our credences in the respective stories. This formula thus helps us to integrate our uncertainty about whether a debunking or non-debunking story holds about our moral intuitions.[23] Even more interestingly, our credences could also be split between different debunking stories – some of which might impact the evidential weight of our intuitions more, and some less. The Jeffrey conditionalization rule can deal with that too, prescribing a final credence that takes all available information into account.

6.4.9 Proximate and Ultimate Causes

Debunking stories could also focus on different kinds of causes that contributed to producing our moral beliefs. Recall that the proximate causes of our moral beliefs operate within our lifetimes (like the immediate psychological mechanisms producing our moral beliefs or

[23] Thanks here to Katie Steele.

intuitions) while their ultimate causes operate outside our lifetimes (such as natural selection operating over many generations). How might these different sources of information contribute to the final moral credence? Recall O'Neill's (2015) argument that when it comes to the reliability of our moral beliefs, information from the proximate causes has priority over information from ultimate causes. This is because if we can tell whether our moral beliefs are reliable just from information about the proximate causes, then information about the ultimate causes tells us nothing more. In the Bayesian framework, we can interpret her as claiming that information about proximate causes *screens off* information about ultimate causes, when it comes to the reliability of our moral beliefs. Suppose now we have two different debunking stories about a specific moral intuition – a *proximate story*, which pertains to proximate causes of that intuition, and an *ultimate story*, which pertains to its ultimate causes. We can read O'Neill as arguing that when we update our moral beliefs, our likelihoods should obey the following two constraints:

$$Cr_1\big(int \,\big|\, proximate\ story\ \&\ ultimate\ story\ \&\ h\big) = Cr_1(int|proximate\ story\ \&\ h)$$

$$Cr_1\big(int \,\big|\, proximate\ story\ \&\ ultimate\ story\ \&\ \sim h\big) = Cr_1(int|proximate\ story\ \&\ \sim h)$$

That is, the likelihood of having a particular intuition, given the truth of a proximate debunking story, an ultimate debunking story, and h, should just be equal to the likelihood given the proximate story and h. (And likewise for the likelihood of having the intuition given that h is false.) This is because the reason why we care about information from the ultimate causes at all is to determine whether the proximate causes are reliable or not – and if we already had a good idea of the reliability of the proximate cause, learning more information from the ultimate causes should produce no further impact. These two constraints caution us against double-counting information about unreliability – when we have such information from both the proximate and ultimate causes, they don't add up linearly.

6.5 Conclusion

In this chapter, I've presented a Bayesian model of the debunking debate. The model involves first updating on the intuition that h, yielding an increased evidential credence. The debunkers then introduce a debunking story as evidence. Updating on both the intuition and the story yields the debunked credence $Cr_3(h)$. This debunked credence is computed using some new likelihoods, including $Cr_1(int|story\&h)$ and $Cr_1(int|story\&\sim h)$. The debunker hopes to constrain these likelihoods, such that the debunked credence is lower than evidential credence. This happens if the debunker can ensure that at least one of the following conditions hold:

$$Cr_1(int \mid story \,\&\, h) < Cr_1(int|h)$$

$$Cr_1(int \mid story \,\&\, \sim h) > Cr_1(int|\sim h)$$

The first condition leads to the epistemic impact of *true positive less likely* – where the true positive rate of our intuitions is revealed to be lower than we initially thought. The second condition leads to the impact of *false positive more likely* – where the false positive rate is revealed to be higher than we thought. I then drew out some important implications: firstly, this debunking effect is only conditional on the agent's having updated erroneously on their intuitions – the debunker shouldn't argue that every agent should reduce confidence. Secondly, the debunker is only licensed to conclude that we should reduce confidence by some degree, and not that we should reduce beyond some threshold. Third, we can quantitatively integrate evidence about the two kinds of epistemic impact, about legitimate background assumptions, about the different possible origins of our moral beliefs, and about the different kinds of causes of our moral beliefs, in order to arrive at a rational credence about morality. I also argued that point for point, the epistemic impact of *false positive more likely* has greater marginal effect on our credence in h than *true positive less likely*. But while there is a clear path from the empirical evidence to *true positive less likely*, the debunkers will have more trouble linking such evidence to *false positive more likely*.

The model I've presented doesn't call the debunking debate in favour of one side or another. Instead, I hope to have clarified some hidden assumptions and drawn out some implications, while being sensitive to the quantitative nature of the debunking argument's epistemic conclusion. With these more clearly in view, we'll be in a better position to decide when, if ever – and by how much – we should change our moral beliefs upon learning about their causal origins.

7 Conclusion

In this thesis, I aimed to investigate the scope and limits of debunking arguments in ethics. I began in Chapter 1 by outlining the structure of a debunking argument, highlighting its different parts and their variations, considering prominent objections, and exploring further issues. A debunking argument starts with some *Empirical Evidence* – this evidence reveals that the target moral beliefs are flawed in some way, through a premise I called *Evidence Reveals Flaw*. In light of this epistemic flaw, the target moral beliefs are undermined, through the *Epistemic Premise* of the debunking argument.

Chapter 2 examined the *Epistemic Premise* of a debunking argument. I argued that putative counterexamples to this premise – beliefs which seem to be flawed but weren't undermined nonetheless – don't work. Some putative counterexamples pertained to an epistemic standard that was inappropriate for moral methodology; other counterexample beliefs just weren't epistemically flawed to begin with. I also argued that we cannot just deny *Epistemic Premise* without a proposing a replacement – since this leads to an unattractive permissiveness towards clearly problematic beliefs. I outlined three strategies for limiting the in-principle scope of a debunking argument by restricting its *Epistemic Premise*, but concluded that all were implausible. I concluded that if we are to resist a debunking argument, it wouldn't be through denying its *Epistemic Premise*.

Chapter 3 interpreted the debunking argument as an undermining relation between our beliefs, and examined three potential objections concerning how debunking arguments might fit into the web of beliefs. The regress objection says roughly that debunking the target judgments might also undermine the basis of the debunking argument, thus disabling its conclusion. I questioned the regress objection at various steps – arguing that Liao et al.'s debunking argument can avoid this objection, and drawing more general lessons about how to resist the regress. I also argued that even if there is a regress, the debunking conclusion still holds. The findings redundancy objection says that the empirical findings are redundant – I responded that at least sometimes, such findings can reveal new information about how our

moral beliefs or intuitions change in response to different factors. The argument redundancy objection says that debunking arguments are evidentially redundant, because they assume what they set out to prove. I argued that as long as debunking arguments don't assume the falsity of the target judgments, they would be evidentially impactful – and that Liao et al.'s argument is one such example.

Chapters 2 and 3 defended the in-principle viability of debunking arguments. In Chapter 4, however, I argued that some debunking arguments still fall short because of poor empirical support. In particular, I argued that current evidence about the evolutionary origins of our moral beliefs doesn't reveal that such beliefs are chancy or inevitable in a way that's epistemically problematic. We don't have evidence of counterfactual evolutionary selves who don't believe that categorical moral reasons exist, and who based their opinions on relevant evidence and sufficiently good evidential processing abilities. We haven't observed scenarios where the moral facts about the existence of categorical moral reasons is changed – so it's unclear how our actual-world evidence reveals the inevitability of our moral beliefs. Moreover, the existence of human anti-realists shows that the belief in categorical moral reasons isn't rendered inevitable by evolutionary influences.

Chapter 5 introduced a novel case study of debunking, drawing on the phenomenon of probability weighting – where we overweight low probabilities, and underweight moderate to high probabilities, in our decision-making. I presented two cases showing how we seem to switch from moral risk seeking to moral risk aversion, when the success probability of a rescue attempt increases. I then considered the normative implications of this pattern – we could a) attempt a debunking argument from the linear weighting of probabilities (which, I argued, would be evidentially redundant), b) attempt a debunking argument from the constant contribution of risk or from an implausible change in risk attitude, or c) attempt a vindicatory argument based on patterns observed in our moral intuitions. I also outlined conditions for when we should endorse a debunking conclusion rather than a vindicatory one, connected this case study with issues raised in previous chapters, and examined potential implications of the debunking conclusion.

Chapter 6 presented a Bayesian model of debunking arguments, which guides us in changing our beliefs in response to debunking. In this model, the empirical evidence constrains our likelihoods, such that the true positive rate of our intuitions is revealed to be lower than we initially thought, or the false positive rate is revealed to be higher than we thought. I argued that the debunking argument should only affect people who updated erroneously on their intuitions before, and that it should only conclude that these people should reduce confidence by some degree (rather than beyond some specific threshold). I also proposed a quantitative method for taking different kinds of evidence into account – concerning the different epistemic flaws at play, the different possible origins of our moral beliefs, and the background normative assumptions we're entitled to make – in order to arrive at a rational credence about morality in light of debunking.

Through my investigations, I've concluded that a debunking argument is successful only if: it impacts the probability that the target moral beliefs are true, rather than their status as justified or knowledge (Chapter 2), it relies on assumptions that are more likely to be true than the target beliefs (and we can use various supporting beliefs and indicators of epistemic status – like level of generality – to show this) (Chapter 3), it does not assume that the target beliefs are false (Chapters 3 and 5), it relies on empirical evidence that reveals new information about how the target beliefs depend on various factors (Chapter 3 and 5), and this evidence can establish that the target beliefs are flawed in the relevant ways (Chapter 4).

You have some moral beliefs. Could you learn anything about the origins of your moral beliefs, or about how these beliefs depend on various causes, that would lead you to conclude that such beliefs are untrustworthy? A qualified 'yes' was the answer I've arrived at. But much work remains to be done to identify these untrustworthy beliefs.

Appendix

In this appendix, I identify the conditions under which the marginal impact of *false positive more likely* is greater than that of *true positive less likely*. First recall the formula for the stage 2 evidential credence:

$$Cr_2(h) = \frac{Cr_1(int|h).Cr_1(h)}{Cr_1(int|h).Cr_1(h) + Cr_1(int|\sim h).Cr_1(\sim h)}$$

For ease of exposition, I'll adopt simpler notation for this formula. Let $Cr_1(int|h) = T$ (for the true positive rate) and $Cr_1(int|\sim h) = F$ (the false positive rate), and let $Cr_1(h) = H$ and $Cr_1(\sim h) = 1 - H$. The formula then becomes

$$Cr_2(h) = \frac{TH}{TH + F(1-H)}$$

The marginal impact of *true positive less likely* can be found by partially differentiating $Cr_2(h)$ with respect to T, and negating the result (because we are interested in a *decrease* in true positive rate).

$$Marginal\ impact\ of\ true\ positive\ less\ likely = -\frac{\partial Cr_2(h)}{\partial T} = -\frac{FH(1-H)}{(TH + F(1-H))^2}$$

With *false positive more likely*, we just partially differentiate $Cr_2(h)$ with respect to F.

$$Marginal\ impact\ of\ false\ positive\ more\ likely$$
$$= \frac{\partial Cr_2(h)}{\partial F}$$
$$= -\frac{TH(1-H)}{(TH+F(1-H))^2}$$

Assume that $T > 0$ and that $0 < H < 1$. *False positive more likely* has a greater marginal impact when it creates a *greater reduction* in $Cr_2(h)$. That is, when

$$Marginal\ impact\ of\ false\ positive\ more\ likely < Marginal\ impact\ of\ true\ positive\ less\ likely$$

$$-\frac{TH(1-H)}{\left(TH+F(1-H)\right)^2} < -\frac{FH(1-H)}{\left(TH+F(1-H)\right)^2}$$

$$\frac{TH(1-H)}{\left(TH+F(1-H)\right)^2} > \frac{FH(1-H)}{\left(TH+F(1-H)\right)^2}$$ **Reverse sign on both sides.**

$$TH(1-H) > FH(1-H)$$ **Multiply by denominator. (Assume T, H > 0)**

$$T > F$$ **Divide by $H(1-H)$. (Assume 0 < H < 1)**

Thus *false positive more likely* has a greater marginal impact if the following conditions hold:

a) the true positive rate is greater than zero, and greater than the false positive rate ($T > F$ and $T > 0$),

b) our prior credence in the hypothesis is between 0 and 1 ($0 < H < 1$).

References

Alfano, M. (2011). Explaining Away Intuitions About Traits: Why Virtue Ethics Seems Plausible (Even if it Isn't). *Review of Philosophy and Psychology*, *2*(1), 121–136. https://doi.org/10.1007/s13164-010-0045-9

Alston, W. P. (1985). Concepts of Epistemic Justification. *The Monist*, *68*(1), 57–89. JSTOR.

Alston, W. P. (2006). *Beyond "Justification": Dimensions of Epistemic Evaluation* (1 edition). Cornell University Press.

Altham, J. E. J. (1983). Ethics of Risk. *Proceedings of the Aristotelian Society*, *84*, 15–29. JSTOR.

Ayala, F. J. (2010). The difference of being human: Morality. *Proceedings of the National Academy of Sciences*, *107*(Supplement 2), 9015–9022. https://doi.org/10.1073/pnas.0914616107

Ballantyne, N. (2012). The Problem of Historical Variability. In D. E. Machuca (Ed.), *Disagreement and Skepticism* (1 edition, pp. 239–259). Routledge.

Ballantyne, N. (2014). Counterfactual Philosophers. *Philosophy and Phenomenological Research*, *88*(2), 368–387. https://doi.org/10.1111/phpr.12068

Barberis, N. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, *27*(1), 173–196. https://doi.org/10.1257/jep.27.1.173

Bedke, M. S. (2009). Intuitive Non-Naturalism meets Cosmic Coincidence. *Pacific Philosophical Quarterly*, *90*(2), 188–209.

Berker, S. (2009). The Normative Insignificance of Neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329. https://doi.org/10.1111/j.1088-4963.2009.01164.x

Berker, S. (2014). Does Evolutionary Psychology Show That Normativity Is Mind-Dependent? In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 215–252). Oxford University Press. http://scholar.harvard.edu/files/sberker/files/berker-evo-psych-mind-dep.pdf

Berker, S. (2015). Coherentism via Graphs. *Philosophical Issues*, *25*(1), 322–352. https://doi.org/10.1111/phis.12052

Bogardus, T. (2016). Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument. *Ethics*, *126*(3), 636–661. https://doi.org/10.1086/684711

Bovens, L., & Hartmann, S. (2004). *Bayesian Epistemology* (1 edition). Oxford University Press.

Brink, D. O. (1989). *Moral Realism and the Foundations of Ethics* (1st edition). Cambridge University Press.

Brosnan, K. (2011). Do the evolutionary origins of our moral beliefs undermine moral knowledge? *Biology & Philosophy*, *26*(1), 51–64. https://doi.org/10.1007/s10539-010-9235-1

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press, USA.

Buchak, L. (2017). Taking Risks Behind the Veil of Ignorance. *Ethics*, *127*(3), 610–644. https://doi.org/10.1086/690070

Carey, B., & Matheson, J. (2013). How Skeptical is the Equal Weight View? In D. E. Machuca (Ed.), *Disagreement and Skepticism* (1 edition, pp. 131–149). Routledge.

Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review*, *116*(2), 187–217.

Christensen, D. (2009). Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass*, *4*(5), 756–767. https://doi.org/10.1111/j.1747-9991.2009.00237.x

Clarke-Doane, J. (2012). Morality and Mathematics: The Evolutionary Challenge. *Ethics*, *122*(2), 313–340. https://doi.org/10.1086/663231

Clarke-Doane, J. (2016). Debunking and Dispensability. In U. D. Leibowitz & N. Sinclair (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford University Press UK.

Climenhaga, N. (2018). Intuitions are Used as Evidence in Philosophy. *Mind*, *127*(505), 69–104. https://doi.org/10.1093/mind/fzw032

Comesaña, J. (2005). Unsafe Knowledge. *Synthese*, *146*(3), 395–404. https://doi.org/10.1007/s11229-004-6213-7

Comesaña, J. (2007). Knowledge and Subjunctive Conditionals. *Philosophy Compass*, *2*(6), 781–791. https://doi.org/10.1111/j.1747-9991.2007.00076.x

Copp, D. (2008). Darwinian Skepticism about Moral Realism. *Philosophical Issues*, *18*(1), 186–206. https://doi.org/10.1111/j.1533-6077.2008.00144.x

Copp, D. (2018). How to avoid begging the question against evolutionary debunking arguments. *Ratio*, 1–15. https://doi.org/10.1111/rati.12222

Cummins, R. C. (1998). Reflection on Reflective Equilibrium. In M. DePaul & W. Ramsey (Eds.), *Rethinking Intuition* (pp. 113–128). Rowman & Littlefield.

Daniels, N. (2015). Can There be Moral Force to Favoring an Identified over a Statistical Life? In I. G. Cohen, N. Daniels, & N. Eyal (Eds.), *Identified versus Statistical Lives: An Interdisciplinary Perspective*. Oxford University Press.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. http://www.gutenberg.org/ebooks/2300

Das, R. (2016). Evolutionary debunking of morality: Epistemological or metaphysical? *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *173*(2), 417–435.

de Lazari-Radek, K., & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics*, *123*(1), 9–31. https://doi.org/10.1086/667837

Dreisbach, S., & Guevara, D. (2019). The Asian Disease Problem and the Ethical Implications Of Prospect Theory. *Noûs*, *53*(3), 613–638. https://doi.org/10.1111/nous.12227

Dworkin, R. (1996). Objectivity and Truth: You'd Better Believe it. *Philosophy & Public Affairs*, *25*(2), 87–139. JSTOR.

Elga, A. (2007). Reflection and Disagreement. *Noûs*, *41*(3), 478–502. https://doi.org/10.1111/j.1468-0068.2007.00656.x

Enoch, D. (2006). Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action. *The Philosophical Review*, *115*(2), 169–198. JSTOR.

Enoch, D. (2010). The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It. *Philosophical Studies*, *148*(3), 413–438.

Fehr-Duda, H., & Epper, T. (2011). Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences. *Annual Review of Economics*, *4*(1), 567–593. https://doi.org/10.1146/annurev-economics-080511-110950

Feldman, R. (2007). Reasonable Religious Disagreements. In L. M. Antony (Ed.), *Philosophers without Gods: Meditations on Atheism and the Secular Life* (1 edition, pp. 194–214). Oxford University Press.

Fennema, H., & Wakker, P. (1997). Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making*, *10*(1), 53–64. https://doi.org/10.1002/(SICI)1099-0771(199703)10:1<53::AID-BDM245>3.0.CO;2-1

FitzPatrick, W. J. (2015). Debunking evolutionary debunking of ethical realism. *Philosophical Studies*, *172*(4), 883–904. https://doi.org/10.1007/s11098-014-0295-y

FitzPatrick, W. J. (2016). Misidentifying the Evolutionary Debunkers' Error: Reply to Mogensen. *Analysis*, *76*(4), 433–437. https://doi.org/10.1093/analys/anw065

Frances, B., & Matheson, J. (2018). Disagreement. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/disagreement/

Fraser, B. J. (2010). Adaptation, Exaptation, By-Products, and Spandrels in Evolutionary Explanations of Morality. *Biological Theory*, *5*(3), 223–227. https://doi.org/10.1162/BIOT_a_00052

Fraser, B. J. (2014). Evolutionary debunking arguments and the reliability of moral cognition. *Philosophical Studies*, *168*(2), 457–473. https://doi.org/10.1007/s11098-013-0140-8

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis*, *23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Goldman, A. (1979). What is Justified Belief? In G. Pappas (Ed.), *Justification and Knowledge* (pp. 1–25). Boston: D. Reidel.

Goldman, A. (2016). Reply to Schaffer. In B. P. McLaughlin & H. Kornblith (Eds.), *Goldman and His Critics* (1 edition, pp. 365–368). Wiley-Blackwell.

Goldman, A., & Beddor, B. (2016). Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/reliabilism/

Gonzalez, R., & Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, *38*(1), 129–166. https://doi.org/10.1006/cogp.1998.0710

Greene, J. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3*. MIT Press.

Handfield, T. (2016). Genealogical Explanations of Chance and Morals. In U. D. Leibowitz & N. Sinclair (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford University Press UK.

Horowitz, T. (1998). Philosophical Intuitions and Psychological Theory. *Ethics*, *108*(2), 367–385. https://doi.org/10.1086/233809

Huemer, M. (2005). *Ethical Intuitionism*. Palgrave Macmillan.

Huemer, M. (2008). Revisionary Intuitionism. *Social Philosophy and Policy*, *25*(01), 368–392. https://doi.org/10.1017/S026505250808014X

Ichikawa, J. J., & Steup, M. (2016). The Analysis of Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/knowledge-analysis/

Isserow, J. (2018). Evolutionary Hypotheses and Moral Skepticism. *Erkenntnis*, 1–21. https://doi.org/10.1007/s10670-018-9993-8

Joyce, R. (2006). *The Evolution of Morality*. MIT Press.

Joyce, R. (2009). The Skeptick's Tale. *Philosophy and Phenomenological Research*, *78*(1), 213–221. https://doi.org/10.1111/j.1933-1592.2008.00241.x

Joyce, R. (2016). Reply: Confessions of a Modest Debunker. In U. D. Leibowitz & N. Sinclair (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford University Press UK.

Kagan, S. (2001). Thinking about Cases. *Social Philosophy and Policy*, *18*(02), 44–63. https://doi.org/10.1017/S0265052500002892

Kahane, G. (2011). Evolutionary Debunking Arguments. *Noûs*, *45*(1), 103–125. https://doi.org/10.1111/j.1468-0068.2010.00770.x

Kahane, G. (2013). The armchair and the trolley: An argument for experimental ethics. *Philosophical Studies*, *162*(2), 421–445. https://doi.org/10.1007/s11098-011-9775-5

Kahane, G. (2016). Is, Ought, and the Brain. In S. M. Liao (Ed.), *Moral Brains: The Neuroscience of Morality* (1 edition, pp. 281–311). Oxford University Press.

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st edition). Farrar, Straus and Giroux.

Kamm, F. M. (1993). *Morality, Mortality: Volume 1: Death and Whom to Save It From*. Oxford University Press USA.

Kamm, F. M. (1998). Moral Intuitions, Cognitive Psychology, and the Harming-Versus-Not-Aiding Distinction. *Ethics*, *108*(3), 463–488.

Kaplan, J. M. (2002). Historical Evidence and Human Adaptations. *Philosophy of Science*, *69*(S3), S294–S304. https://doi.org/10.1086/341853

King, N. L. (2012). Disagreement: What's the Problem? or A Good Peer is Hard to Find. *Philosophy and Phenomenological Research*, *85*(2), 249–272. https://doi.org/10.1111/j.1933-1592.2010.00441.x

Klenk, M. (2018). Evolution and Moral Disagreement. *Journal of Ethics and Social Philosophy*, *14*(2). https://doi.org/10.26556/jesp.v14i2.476

Knobe, J. (2019). Philosophical Intuitions Are Surprisingly Robust Across Demographic Differences. *Epistemology and Philosophy of Science*, *56*, 29–36.

Kotzen, M. (forthcoming). A Formal Account of Epistemic Defeat. *Synthese Library*. http://matthewkotzen.net/matthewkotzen.net/Research_files/Klein.pdf

Kühberger, A. (1995). The Framing of Decisions: A New Look at Old Problems. *Organizational Behavior and Human Decision Processes*, *62*(2), 230–240. https://doi.org/10.1006/obhd.1995.1046

Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology*, *25*(3), 311–330. https://doi.org/10.1080/09515089.2012.660140

Lahti, D. C. (2003). Parting with illusions in evolutionary ethics. *Biology and Philosophy*, *18*(5), 639–651. https://doi.org/10.1023/A:1026356412003

Lange, M. (1999). Calibration and the Epistemological Role of Bayesian Conditionalization. *The Journal of Philosophy*, *96*(6), 294–324. JSTOR. https://doi.org/10.2307/2564680

Lazar, S. (forthcoming). Duty and Doubt. *Journal of Practical Ethics*.

Lazar, S. (2018). Limited Aggregation and Risk. *Philosophy & Public Affairs*, *46*(2), 117–159. https://doi.org/10.1111/papa.12115

Lazar, S. (2019). Risky Killing: How Risks Worsen Violations of Objective Rights. *Journal of Moral Philosophy*, *16*(1), 1–26.

Lewontin, R. C. (1998). The evolution of cognition: Questions we will never answer. *An Invitation to Cognitive Science*, *4*, 107–132.

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661–671. https://doi.org/10.1080/09515089.2011.627536

Lillehammer, H. (2010). Methods of ethics and the descent of man: Darwin and Sidgwick on ethics and evolution. *Biology & Philosophy*, *25*(3), 361–378. https://doi.org/10.1007/s10539-010-9204-8

Lutz, M. (2018). What Makes Evolution a Defeater? *Erkenntnis*, *83*(6), 1105–1126. https://doi.org/10.1007/s10670-017-9931-1

Machery, E. (2017). *Philosophy Within Its Proper Bounds* (1 edition). Oxford University Press.

Machery, E., & Mallon, R. (2010). Evolution of Morality. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (p. 3). Oxford University Press.

Mackie, J. (1977). *Ethics: Inventing right and wrong*. Penguin UK.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology. General*, *143*(3), 1185–1198. https://doi.org/10.1037/a0034207

Matheson, J. (2015). *Disagreement and Epistemic Peers*. https://doi.org/10.1093/oxfordhb/9780199935314.013.13

McGrath, S. (2008). Moral Disagreement and Moral Expertise. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics: Volume 4* (pp. 87–108). Oxford University Press.

McGrath, S. (2014). Relax? Don't Do It! Why Moral Realism Won't Come Cheap. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics: Volume 9* (1 edition, pp. 186–213). Oxford University Press.

McPherson, T. (2014). A Case for Ethical Veganism. *Journal of Moral Philosophy*, *11*(6), 677–703. https://doi.org/10.1163/17455243-4681041

Mogensen, A. L. (2015). Evolutionary debunking arguments and the proximate/ultimate distinction. *Analysis*, *75*(2), 196–203. https://doi.org/10.1093/analys/anv013

Mogensen, A. L. (2016a). Do evolutionary debunking arguments rest on a mistake about evolutionary explanations? *Philosophical Studies*, *173*(7), 1799–1817. https://doi.org/10.1007/s11098-015-0579-x

Mogensen, A. L. (2016b). Contingency Anxiety and the Epistemology of Disagreement. *Pacific Philosophical Quarterly*, *97*(4), 590–611. https://doi.org/10.1111/papq.12099

Morton, J. (2016). A New Evolutionary Debunking Argument Against Moral Realism. *Journal of the American Philosophical Association*, *2*(2), 233–253. https://doi.org/10.1017/apa.2016.14

Nichols, S. (2014). Process Debunking and Ethics. *Ethics*, *124*(4), 727–749. https://doi.org/10.1086/675877

Nozick, R. (1974). *Anarchy, State, and Utopia* (First Edition edition). Basic Books.

Nozick, R. (1981). *Philosophical Explanations*. Harvard University Press.

Oddie, G., & Milne, P. (1991). Act and value: Expectation and the representability of moral theories. *Theoria*, *57*(1–2), 42–76. https://doi.org/10.1111/j.1755-2567.1991.tb00540.x

O'Neill, E. (2015). Which Causes of Moral Beliefs Matter? *Philosophy of Science*, *82*(5), 1070–1080. https://doi.org/10.1086/683441

Otsuka, M. (2015). Risking Life and Limb: How to Discount Harms by Their Improbability. In I. G. Cohen, N. Daniels, & N. Eyal (Eds.), *Identified versus Statistical Lives: An Interdisciplinary Perspective*. Oxford University Press.

Parfit, D. (1986). *Reasons and persons*. OUP Oxford.

Pollock, J. L. (1970). The Structure of Epistemic Justification. In N. Rescher (Ed.), *Studies in the Theory of Knowledge*. Basil Blackwell.

Pust, J. (2019). Intuition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/intuition/

Quiggin, J. (1993). *Generalized Expected Utility Theory: The Rank-Dependent Model*. Springer Netherlands. https://doi.org/10.1007/978-94-011-2182-8

Rabinowitz, D. (2018). The Safety Condition for Knowledge. In *Internet Encyclopedia of Philosophy*. https://www.iep.utm.edu/safety-c/

Rasmussen, K. B. (2012). Should the probabilities count? *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *159*(2), 205–218. JSTOR.

Rawls, J. (1999). *A Theory of Justice* (Revised edition). Belknap Press.

Rini, R. A. (2016). Debunking Debunking: A Regress Challenge for Psychological Threats to Moral Judgment. *Philosophical Studies*, *173*(3), 675–697.

Rini, R. A. (2017). Why moral psychology is disturbing. *Philosophical Studies*, *174*(6), 1439–1458. https://doi.org/10.1007/s11098-016-0766-4

Roush, S. (2007). *Tracking Truth: Knowledge, Evidence, and Science* (1 edition). Oxford University Press.

Ruse, M., & Wilson, E. O. (1985). The evolution of ethics. *New Scientist*, *108*(1478), 50–52.

Ruse, M., & Wilson, E. O. (1986). Moral Philosophy as Applied Science. *Philosophy*, *61*(236), 173–192.

Schafer, K. (2010). Evolution and Normative Scepticism. *Australasian Journal of Philosophy*, *88*(3), 471–488. https://doi.org/10.1080/00048400903114219

Schechter, J. (2017). Explanatory Challenges in Metaethics. In T. McPherson & D. Plunkett (Eds.), *The Routledge Handbook of Metaethics* (1 edition, pp. 443–458). Routledge.

Shafer-Landau, R. (2012). Evolutionary Debunking, Moral Realism, and Moral Knowledge. *Journal of Ethics and Social Philosophy*, *7*(1).

Shogenji, T. (2012). Internalism and Externalism in Meliorative Epistemology. *Erkenntnis*, *76*(1), 59–72. https://doi.org/10.1007/s10670-011-9323-x

Sidgwick, H. (1907). *The Methods of Ethics* (7th ed.). Macmillan. https://www.gutenberg.org/files/46743/46743-h/46743-h.htm

Sinclair, N. (2018). Belief Pills and the Possibility of Moral Epistemology. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics 13*. Oxford University Press.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics*, *9*(3–4), 331–352. https://doi.org/10.1007/s10892-005-3508-y

Sinnott-Armstrong, Walter. (2006). *Moral Skepticisms* (1 edition). Oxford University Press.

Sinnott-Armstrong, Walter. (2007). Framing Moral Intuitions. In *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity* (pp. 47–76). A Bradford Book.

Sinnott-Armstrong, Walter. (2011). Emotion and Reliability in Moral Psychology. *Emotion Review*, *3*(3), 288–289. https://doi.org/10.1177/1754073911402382

Skarsaune, K. O. (2011). Darwin and moral realism: Survival of the iffiest. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *152*(2), 229–243. JSTOR.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press.

Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives*, *13*, 141–153. JSTOR.

Sosa, E. (2000). Skepticism and Contextualism. *Philosophical Issues*, *10*, 1–18. JSTOR.

Spector, H. (2016). Decisional Nonconsequentialism and the Risk Sensitivity of Obligation. *Social Philosophy and Policy*, *32*(2), 91–128. https://doi.org/10.1017/S0265052516000121

Srinivasan, A. (2015). The Archimedean Urge. *Philosophical Perspectives*, *29*(1), 325–362. https://doi.org/10.1111/phpe.12068

Sterelny, K., & Fraser, B. (2017). Evolution and Moral Realism. *The British Journal for the Philosophy of Science*, *68*(4), 981–1006. https://doi.org/10.1093/bjps/axv060

Steup, M. (2018). Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/epistemology/

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, *127*(1), 109–166. https://doi.org/10.1007/s11098-005-1726-6

Street, S. (2008). Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About. *Philosophical Issues*, *18*(1), 207–228. https://doi.org/10.1111/j.1533-6077.2008.00145.x

Street, S. (2009). Evolution and the Normativity of Epistemic Reasons. *Canadian Journal of Philosophy Supplementary Volume*, *35*, 213–248. https://doi.org/10.1080/00455091.2009.10717649

Sudduth, M. (2019). *Defeaters in Epistemology*. Internet Encyclopedia of Philosophy. https://www.iep.utm.edu/ep-defea/#SH1b

Talbott, W. (2016). Bayesian Epistemology Supplement—Probability Laws. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/epistemology-bayesian/supplement1.html

Taurek, J. M. (1977). Should the Numbers Count? *Philosophy & Public Affairs*, *6*(4), 293–316. JSTOR.

Tersman, F. (2008). The reliability of moral intuitions: A challenge from neuroscience. *Australasian Journal of Philosophy*, *86*(3), 389–405. https://doi.org/10.1080/00048400802002010

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Tropman, E. (2014). Evolutionary debunking arguments: Moral realism, constructivism, and explaining moral knowledge. *Philosophical Explorations*, *17*(2), 126–140. https://doi.org/10.1080/13869795.2013.855807

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Van Roojen, M. (1999). Reflective moral equilibrium and psychological theory. *Ethics*, *109*(4), 846–857.

Vavova, K. (forthcoming). The limits of rational belief revision: A dilemma for the Darwinian debunker. *Noûs*. https://doi.org/10.1111/nous.12327

Vavova, K. (2014). Debunking Evolutionary Debunking. *Oxford Studies in Metaethics*, *9*, 76–101.

Vavova, K. (2015). Evolutionary Debunking of Moral Realism. *Philosophy Compass*, *10*(2), 104–116. https://doi.org/10.1111/phc3.12194

Walden, K. (2014). The Aid That Leaves Something to Chance. *Ethics*, *124*(2), 231–241. JSTOR. https://doi.org/10.1086/673438

Weisberg, J. (2017). Formal Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/formal-epistemology/

White, R. (2010). You Just Believe That Because…. *Philosophical Perspectives*, *24*(1), 573–615.

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995

Wielenberg, E. J. (2010). On the Evolutionary Debunking of Morality. *Ethics*, *120*(3), 441–464.

Wielenberg, E. J. (2014). *Robust Ethics: The Metaphysics and Epistemology of Godless Normative Realism* (1 edition). Oxford University Press.

Wright, C. (1991). Scepticism and Dreaming: Imploding the Demon. *Mind*, *100*(1), 87–116. JSTOR.